Marginal cure rate models for long-term survivors

by

Jianfeng Chen

M.S., Beijing Jiaotong University, China, 2009

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

Two-component mixture models for long-term survivors, known as cure rate models, have been widely used and intensively discussed in the literature. In most applications, much of attention has been put on interpreting the covariate effects on the two components of the model: the cure fraction and the conditional hazard rate. However, for this mixture model, it is very challenging to give a straightforward interpretation of covariate effects on the overall survival responses, especially when the covariates are shared by these two components of the model. By overall survival responses, we mean the population survival outcomes such as the overall survival rate or the overall instantaneous death rate.

In our study, we propose two marginal cure rate models that can offer a general framework to investigate the covariate effects on the overall survival outcomes from the marginal perspective and, most importantly, provide nice interpretations. These two models are named as Marginal Mean Survival Rate Model and Marginal Mean Hazard Rate Model. Technically, novel reparameterizations are used to relate the covariates directly to the marginal mean survival rate or hazard rate. These parameterizations then can be purposely imposed into the likelihood function of a standard cure rate model and all parameters can be estimated via the regular likelihood approach. We evaluate the proposed marginal models extensively with simulation studies and further use the liver cancer data from the SEER registry as an illustration of the proposed model. Moreover, we propose a semi-parametric approach based on the Bernstein polynomials to relax the assumption of parametric baseline hazard for the noncured subpopulation. The performance of the proposed semi-parametric method is also evaluated through an extensive simulation study and illustrated with SEER liver cancer data.

Finally, as motivated by the microarray data of breast cancer patients from The Cancer Genome Atlas (TCGA) program, we extend the proposed marginal mean hazard rate model

to high-dimensional settings. We handle the high-dimensional covariates with the use of variable selection method based on LASSO-type penalized likelihood function. The model estimation can be easily done with a minimum programming effort by using the techniques of quadratic approximation and cyclic coordinate descent algorithm. The simulation results show that our approach for high-dimensional settings performs reasonably well in terms of low False Positive and False Negative Rates. We then apply our approach to a subset of TCGA microarray data for illustration.

Marginal cure rate models for long-term survivors

by

Jianfeng Chen

M.S., Beijing Jiaotong University, China, 2009

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Wei-Wen Hsu

Copyright

Abstract

Two-component mixture models for long-term survivors, known as cure rate models, have been widely used and intensively discussed in the literature. In most applications, much of attention has been put on interpreting the covariate effects on the two components of the model: the cure fraction and the conditional hazard rate. However, for this mixture model, it is very challenging to give a straightforward interpretation of covariate effects on the overall survival responses, especially when the covariates are shared by these two components of the model. By overall survival responses, we mean the population survival outcomes such as the overall survival rate or the overall instantaneous death rate.

In our study, we propose two marginal cure rate models that can offer a general framework to investigate the covariate effects on the overall survival outcomes from the marginal perspective and, most importantly, provide nice interpretations. These two models are named as Marginal Mean Survival Rate Model and Marginal Mean Hazard Rate Model. Technically, novel reparameterizations are used to relate the covariates directly to the marginal mean survival rate or hazard rate. These parameterizations then can be purposely imposed into the likelihood function of a standard cure rate model and all parameters can be estimated via the regular likelihood approach. We evaluate the proposed marginal models extensively with simulation studies and further use the liver cancer data from the SEER registry as an illustration of the proposed model. Moreover, we propose a semi-parametric approach based on the Bernstein polynomials to relax the assumption of parametric baseline hazard for the noncured subpopulation. The performance of the proposed semi-parametric method is also evaluated through an extensive simulation study and illustrated with SEER liver cancer data.

Finally, as motivated by the microarray data of breast cancer patients from The Cancer Genome Atlas (TCGA) program, we extend the proposed marginal mean hazard rate model

to high-dimensional settings. We handle the high-dimensional covariates with the use of variable selection method based on LASSO-type penalized likelihood function. The model estimation can be easily done with a minimum programming effort by using the techniques of quadratic approximation and cyclic coordinate descent algorithm. The simulation results show that our approach for high-dimensional settings performs reasonably well in terms of low False Positive and False Negative Rates. We then apply our approach to a subset of TCGA microarray data for illustration.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I want to take this opportunity to express my sincere gratitude and appreciation to my advisor Dr. Wei-Wen Hsu for the continuous support over five years' Ph.D. study and research, for his patience, motivation, and immense knowledge. He has provided not only the research topics, but also invaluable advice and guidance that have made me a better person. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my supervisory committee: Dr. James Neill, Dr. Weixing Song, Dr. Gyuhyeong Goh, and Dr. Shing I. Chang. They positively impact on me not only for their insightful comments and encouragement, but also for the questions which incented me to widen my research from various perspectives.

Last but not least, I would like to thank my family: my wife, my parents, and my lovely kid for supporting me spiritually throughout writing this dissertation and my life in general.

Chapter 1

Introduction

In many clinical studies, survival analysis is a critical method to analyze the data where the response of interest is time to the occurrence of event, such as death or relapse of disease. In order to evaluate the covariates effects on the survival outcomes, researchers often use the survival models based on regression techniques such as proportional hazard (PH) model in which the covariates effects are typically interpreted for the hazard rate (Prentice and Kalbfleisch, 1979). However, this type of survival models may not work for the data with the presence of long-term survivors. Such data often exhibit heavy censoring with these long-term survivors, suggesting a proportion of cure in the population. For this reason, cure rate models that can incorporate a cure fraction are commonly suggested and widely used in the literature for analyzing such data (for example, see Sposto, 2002; Bejan-Angoulvant et al., 2008; Cucchetti et al., 2015). A cure rate model is typically a two-component mixture model which assumes the overall population is a mixture of cured and uncured subpopulations. This class of model has been well studied for both parametric and nonparametric approaches (see, Boag, 1949; Berkson and Gage, 1952; Farewell, 1982, 1986; Taylor, 1995; Sy and Taylor, 2000). However, the way to interpret the covariate effects under such model is pretty restricted since we are only allowed to make interpretations on the cure fraction and the baseline hazard rate from the conditional perspective. Thus, it becomes very challenging for studies that are attempting to directly interpret the covariate

effects on the marginal survival rate or hazard rate.

To interpret the covariate effects from a marginal perspective has been studied for the class of mixture models in some papers. For example, Heagerty et al. (2000) reparametrized the marginal mean for the multilevel model in which covariates are regressed on the marginal mean directly, rather than on the conditional mean given the latency. Albert et al. (2014) employed an Average Predicted Value method for zero-inflated models to evaluate the overall exposure effect by comparing the marginal mean counts of exposure and non-exposure group after integrating over all other extraneous covariates. Long et al. (2014, 2015), Smith et al. (2014), and Todem et al. (2016) employed marginalized methods relating the marginal mean to covariates directly in order to offer a straightforward inference about the overall population mean. Other related applications for this idea can be found in Wang and Griswold (2016, 2017), where marginal technique is adopted to give a natural explanation of covariates effects for the Tobit regression model. However, there is no study, to our best knowledge, in the literature to discuss the marginal model under the framework of cure rate model.

In this dissertation, we develop marginal cure rate models with novel parameterizations to relate covariates directly to the marginal mean survival rate or marginal mean hazard rate. The proposed models yield a feasible solution to obtain the interpretation of the covariate effects on the marginal survival outcomes. Our models based on these novel parameterizations can be easily fitted by using routine statistical software such as SAS NLMIXED procedure with a minimum programming effort. Under the framework of marginal cure rate model, we firstly propose a parametric method and assume a weibull baseline hazard for the uncured subpopulation. Then we employ the Bernstein polynomials to relax the restriction of Weibull baseline hazard for the uncured latency. This semi-parametric technique would extend the usage of our proposed marginal model.

As motivated by the microarray data of breast cancer patients from The Cancer Genome Atlas (TCGA), we further extend one of our marginal cure rate models to high-dimensional data in the sense of a massive number of covariates (i.e., microarrays). Coupled with the use of LASSO-type penalized likelihood function, the extended model can successfully identify the critical microarrays that are significantly related to the overall survival outcome and

provide a straightforward interpretation.

The rest of the dissertation is organized as follows: In Chapter 2, we briefly introduce the standard cure rate model and discuss the challenges of making interpretation for the covariate effects on the overall population. In Chapter 3, we propose a naïve marginal mean survival rate model which can relate covariates to the overall mean survival rate via a novel parameterization. In Chapter 4, we propose a marginal mean hazard rate model in which the covariates are regressed on the marginal mean hazard rate. This method is then evaluated through simulation studies and illustrated with liver cancer data from the SEER registry. In chapter 5, we employ a semi-parametric technique to relax the assumption of Weibull baseline hazard for the uncured latency. More specifically, Bernstein polynomials are employed for the baseline hazard, and simulation results indicate the consistency of estimators obtained from this new method. In Chapter 6, we extend the marginal mean hazard rate model to high-dimensional data coupled with the use of variable selection based on LASSO-type penalized likelihood function. Simulations are conducted, and the application to TCGA breast cancer data is used to illustrate the proposed marginal model in high-dimensional settings.

Chapter 2

Two-component mixure cure rate model

It is getting common to see the survival data contain two types of subjects. One type of subjects is susceptible to the event of interest and the other type is not, which means that we may observe some long-term survivors who will never experience the event. This type of survival data often shows a non-negligible proportion of censoring after a long follow-up and its associated survival curve levels off at a specific survival rate. It is reasonable to assume the population is a mixture of uncured subpopulation consisting of the susceptible subjects and cured subpopulation consisting of all the long-term survivors. However, standard survival models are not able to accommodate these two different types of subjects. As a solution, an extended survival model called two-component mixture cure rate model (Boag, 1949; Farewell, 1982) is proposed to model such survival data from the heterogeneous population.

2.1 Two-component mixture cure rate model

Assuming the individual i $(i = 1, 2, ..., n)$ is from uncured group that will eventually experience the event $(U_i = 1)$ or cured group that will never experience the event $(U_i = 0)$, with $P(U_i = 1) = \pi_i$. This probability π_i is also referred to as uncure fraction. The indicator

U is called the membership latent variable which is partially observed. Let t_i denote the time to the event or the censoring time, and $S_u(t_i)$ be the survival function for the uncured individuals. The marginal survival function for the overall population is defined as

$$S_M(t_i) = (1 - \pi_i) + \pi_i S_u(t_i), \quad 0 \leq \pi_i \leq 1 \tag{2.1}$$

In practice, the uncure fraction π_i is often related to covariates $\boldsymbol{z}_i$ through a logit link function (Farewell, 1982), that is, $\text{logit}(\pi_i) = \boldsymbol{\xi' z_i}$. Unless $\pi_i = 1$, the marginal survival function $S_M(t_i)$ is improper in the sense that it has the range of $[1 - \pi_i, 1]$. The conditional survival function $S_u(t_i)$ for the uncured individuals could be parametric (Boag, 1949; Farewell, 1982; Fan et al., 2017) or semi-parametric (Cox, 1972; Sy and Taylor, 2000; Wang et al., 2012) under the proportional hazard assumption. The hazard rate $h_u(t_i)$ for the uncured individual with the survival time t_i is

$$h_u(t_i|\boldsymbol{w}_i) = h_{u0}(t_i) \exp\left\{\boldsymbol{\eta' w_i}\right\}, \quad t_i \geq 0,$$

where $\boldsymbol{w_i}$ is the covariate vector for the i^{th} individual. The parameter $\boldsymbol{\eta}$ is a vector of unknown regression coefficients. Cox (1972) assumed an arbitrary nuissance baseline hazard $h_{u0}(t_i)$ and used a partial likelihood function that only invovles $\boldsymbol{\eta}$ for further parameter estimation. However, this technique does not work for the cure rate model since $h_{u0}(t_i)$ cannot be cancelled out in the likelihood function. If the baseline hazard function is assumed parametrically, then the η can be estimated via a regular maximum likelihood approach. Some parametric baseline hazard functions suggested in the literature are including lognormal (Boag, 1949), exponential (Boag, 1949; Berkson and Gage, 1952; Fan et al., 2017) and Weibull (Farewell, 1982, 1986; Hsu et al., 2016). The most popular baseline hazard is Weibull since it is more flexible compared to exponential due to its additional scale parameter. The Weibull baseline hazard is

$$h_{u_0}(t_i) = \alpha \lambda t_i^{\alpha - 1},$$

where $\alpha > 0$ and $\lambda > 0$ are the scale and shape parameters. For the uncured subpopulation,

the survival function in Equation 2.1 could be derived by the cumulative hazard rate function $H_u(t_i) = \int_0^\infty h_u(v_i)dv_i$, which is

$$\begin{aligned} S_u(t_i|\boldsymbol{w}_i) &= e^{-H_u(t_i)} \\ &= \left[S_{u_0}(t_i)\right]^{\exp\{\boldsymbol{\eta'}\boldsymbol{w_i}\}}, \end{aligned} \tag{2.2}$$

where S_{u0} is the baseline survival function. When Weibull distribution is assumed, the baseline survival function is $S_{u0} = \exp\{-\lambda t_i^\alpha\}$.

In real studies, it is also important to justify that the cure rate model is appropriate for data fitting. Zhao et al. (2009) developed a score test procedure to determine whether the cure fraction is significant. Using above notations, we are interested in testing $H_0 : \phi_i = 0$ against $H_1 : \phi_i > 0$, where $\phi_i = (1 - \pi_i)/\pi_i$, then the score test statistic could be written as

$$S(\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, 0) = U'(\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, 0)\hat{\Gamma}^{-1}U(\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, 0)$$

where $\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\eta}}$ are parameter estimates under the null hypothesis $\phi_i = 0$. The U is corresponding score function and $\hat{\Gamma}$ is the Fisher information matrix under the alternative hypothesis but evaluated at $(\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\eta}}, 0)$, which are estimates under the null hypothesis.

Hsu et al. (2016) further constructed a sup-score test statistic by using empirical process, which could take advantage of covariate information to assess the cure fraction.

2.2 Likelihood function for the two-component cure rate model

Suppose the survival data for the i^{th} individual is $\{t_i, \delta_i, \boldsymbol{z_i}, \boldsymbol{w_i}\}$, where t_i is the survival time, and δ_i is the censoring indicator ($\delta_i = 1$, noncensored; $\delta_i = 0$, censored). The covariates $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$ are related to the uncure rate π_i and the hazard rate $h_u(t_i)$, respectively. Assuming the censoring mechanism is noninfomative and independent with U_i, then the likelihood

function for the cure rate model is

$$\begin{aligned}\mathcal{L}(\alpha, \lambda, \boldsymbol{\eta}, \boldsymbol{\xi}|\boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{W}, \boldsymbol{Z}) &= \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \Big[\pi_i f_u(t_i)\Big]^{\delta_i} \Big[1 - \pi_i + \pi_i S_u(t_i)\Big]^{1-\delta_i},\end{aligned} \tag{2.3}$$

where the conditional survival function $S_u(t_i)$ is given by Equation 2.2 if proprotional hazard is assumed for the uncured individuals. Under the same assumption, the $f_u(t_i)$ has the following form:

$$\begin{aligned}f_u(t_i) &= h_u(t_i) S_u(t_i) \\ &= h_{u_0}(t_i) \exp\{\boldsymbol{\eta'} \boldsymbol{w_i}\} \big[S_{u_0}(t_i)\big]^{\exp\{\boldsymbol{\eta'} \boldsymbol{w_i}\}}\end{aligned} \tag{2.4}$$

The parameter estimation can be done by maximizing the log-likelihood function using Newton-Ralphon or quasi-Newton method in most statistical software such as SAS NLMIXED procedure. Other techniques such as EM algorithm (Peng et al., 1998; Li and Taylor, 2002; Wang et al., 2012) could also be used for the parameter estimation.

2.3 Difficulty in interpretating from the marginal perspective

In practice, it is often of interest to evaluate the covariate effects in this cure rate model. The covariate effects on the response of uncured subpopulation can be evaluated by the coefficient vector $\boldsymbol{\eta}$ and the covariate effects on the cure fraction could be easily evaluated by $\boldsymbol{\xi}$ via the logit link function. However, it is very challenging to interpret the covariate effects on the marginal responses since marginal response $S_M(t_i)$ or $h_M(t_i)$ are not simple functions of parameters ξ and η under the framework of two-component mixture models. For example, assuming $\pi_i = e^{\boldsymbol{\xi' z_i}}/(1 + e^{\boldsymbol{\xi' z_i}})$ and $h_u(t_i)$ is the same as Equation 2.1, then

marginal hazard rate for the two-component mixture cure rate model is

$$\begin{aligned} h_M(t) &= \frac{\pi_i f_u(t_i)}{1-\pi_i+\pi_i S_u(t_i)} \\ &= \frac{\exp\{\boldsymbol{\xi'} \boldsymbol{z_i}\} h_u(t_i)}{\exp\{\boldsymbol{\xi'} \boldsymbol{z_i}\} + S_u(t_i)^{-1}} \quad \text{by Equation (2.4)} \\ &= \frac{\exp\{\boldsymbol{\xi'} \boldsymbol{z_i} + \boldsymbol{\eta'} \boldsymbol{w_i}\} h_{u0}(t_i)}{\exp\{\boldsymbol{\xi'} \boldsymbol{z_i}\} + \left[S_{u_0}(t_i)\right]^{-\exp\{\boldsymbol{\eta'} \boldsymbol{w_i}\}}}, \quad \text{for } t > 0 \end{aligned} \tag{2.5}$$

Clearly, the marginal hazard function $h_M(t_i)$ no longer satisfies the proportional hazard rate assumption, which makes the interpretation difficult from the marginal perspective. Specifically, for two individuals with different covariates $\boldsymbol{z}$ and $\boldsymbol{z^*}$, the hazard ratio is

$$\frac{h_M(t_i|\boldsymbol{z})}{h_M(t_i|\boldsymbol{z^*})} = \exp\left\{(\boldsymbol{\xi}+\boldsymbol{\eta})'(\boldsymbol{z}-\boldsymbol{z^*})\right\} \frac{\exp\{\boldsymbol{\xi'} \boldsymbol{z^*}\} + \left[S_{u_0}(t)\right]^{-\exp\{\boldsymbol{\eta'} \boldsymbol{z^*}\}}}{\exp\{\boldsymbol{\xi'} \boldsymbol{z}\} + \left[S_{u_0}(t)\right]^{-\exp\{\boldsymbol{\eta'} \boldsymbol{z}\}}} \tag{2.6}$$

This hazard ratio is clearly not a constant over time and the baseline hazard $h_{u0}(t_i)$ needs to be specified if a closed form of the ratio is desired. Due to these difficulties, a new method to interpret the covariate effects on the overall population is much needed.

Chapter 3

Marginal mean survival rate model

The two-component cure rate model is appealing to accommodate time-to-event data when long-term survivors are present in the data. With the regression technique, the two-component cure rate model can provide the interpretation of covariate effects on the cure fraction and the uncured subpopulation survival easily. However, for the studies that are focusing on the covariate effects on the overall survival rate or hazard rate, two-component cure rate model is not appropriate for such use. It is known that $S_M(t_i) = (1 - \pi_i) + \pi_i S_u(t_i)$, and the overall survival function $S_M(t_i)$ is confounded by the subpopulation survival rate $S_u(t_i)$ and cure fraction π_i, where the $S_u(t_i)$ and π_i are related to their own covariates separately. This makes the interpretation of covariate effects for the marginal survival function extremely challenging. Likewise, covariate effects on the marginal hazard rate $h_M(t_i)$ is also diffcult to interpret for the same reason.

In this chapter, we employ the idea of marginal methods (Heagerty et al., 2000; Smith et al., 2014; Long et al., 2014, 2015; Todem et al., 2016) that relate the marginal mean to covariates directly for which we could interpret the covariate effects easily from the marginal prospective. Existing marginal methods focus on zero-inflated models or generalized linear models with random effects. However, it is challenging to apply the same idea directly due to the fact that the marginal survival function in the cure rate model is not standard.

For this, we propose two methods which are marginal mean survival rate model (MMSR)

and marginal mean hazard rate model (MMHR, this will be introduced in Chapter 4). Techinically, we relate covariates to the overall mean survival rate and the overall mean hazard rate through novel link functions. In this chapter, we mainly focus on the MMSR model.

3.1 Formulation of marginal mean survival rate model

The proposed MMSR model can be used to evaluate the covariate effects on the overall mean survival rate $E[S_M(t_i)]$. In this model, covariates are related to $E[S_M(t_i)]$ by a novel link function which has the same range as $E[S_M(t_i)]$. As t_i is a continuous random variable, it is clear that $S_u(t_i)$ is also a random variable. Then we could take the expectation with respect to t_i on the both sides of Equation 2.1,

$$E[S_M(t_i)] = 1 - \pi_i + \pi_i E[S_u(t_i)]$$

As known, $0 \leq S_u(t_i) \leq 1$ and $S_u(t_i) = 1 - F_u(t_i)$, where $F_u(t_i)$ is the cumulative density function for t_i. By the probability integral transformation theory and the assumption of continuity of t_i, we have $F_u(t_i) \sim U(0,1)$. As a result, the $S_u(t_i) \sim U(0,1)$ and $E[S_u(t_i)] = 1/2$. Therefore,

$$E[S_M(t_i)] = 1 - \frac{1}{2}\pi_i \tag{3.1}$$

It is clear that the range of $E[S_M(t_i)]$ is $[0, \frac{1}{2}]$ for $0 \leq \pi_i \leq 1$. We propose a novel link function to relate $E[S_M(t_i)]$ to covariates and given as,

$$E[S_M(t_i)] = \frac{1 + e^{\boldsymbol{\gamma}' \boldsymbol{x_i}}}{1 + 2e^{\boldsymbol{\gamma}' \boldsymbol{x_i}}}, \tag{3.2}$$

where $\boldsymbol{\gamma}$ is the marginal parameter vector and this link function has the same support as of $E[S_M(t_i)]$. This link function provides a direct interpretation of covariate effects on the marginal mean survival rate. To estimate the marginal coefficients $\boldsymbol{\gamma}$, we borrow the framework of standard cure rate model and replace π_i with a function of $E[S_M(t_i)]$ in the

likelihood function (Equation 2.3). By combing Equations 3.1 and 3.2, we can have

$$\pi_i = 2 - 2E[S_M(t_i)] \quad \text{or} \quad \pi_i = \frac{e^{\boldsymbol{\gamma' x_i}}}{0.5 + e^{\boldsymbol{\gamma' x_i}}} \tag{3.3}$$

It is interesting to see that the uncure fraction π_i under the marginal mean survival rate model is very similar to the π_i under the conditional cure rate model where $\pi_i = e^{\boldsymbol{\xi' z_i}}/(1 + e^{\boldsymbol{\xi' z_i}})$. By assuming $\boldsymbol{x_i} = \boldsymbol{z_i}$, then the only difference lies in the first constant in both denominators, meaning that the interpretation for the covariate effects on the marginal mean survival rate is similar to the interpretation for the cure fraction under the regular cure rate model.

3.2 Likelihood function and estimation

Suppose we have independently observed survival data $(t_i, \delta_i, \boldsymbol{x_i}, \boldsymbol{w_i})$, where t_i is the observed event time, δ_i is the censoring indicator ($\delta_i = 1$, noncensored; $\delta_i = 0$, censored). Covariate vectors $\boldsymbol{x_i}$ and $\boldsymbol{w_i}$ are the covariates related to marginal mean survival rate $E[S_M(t_i)]$ and hazard rate $h_u(t_i)$, respectively. Then, likelihood function for the marginal mean survival rate model under the independence assumption is

$$\begin{aligned}\mathcal{L}(\alpha, \lambda, \boldsymbol{\gamma}, \boldsymbol{\eta} | \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{W}) &= \prod_{i=1}^{n} f_M(t_i)^{\delta_i} \prod_{i=1}^{n} S_M(t_i)^{1-\delta_i} \\ &= \prod_{\delta_i=1} \Big\{\pi_i f_u(t_i)\Big\} \prod_{\delta_i=0} \Big\{2E[S_M(t_i)] - 1 + \big\{2 - 2E[S_M(t_i)]\big\} S_u(t_i)\Big\}\end{aligned}$$

In our model, we assume a proportional hazard model $h_u(t_i) = h_{u0}(t_i) \exp\{\boldsymbol{\eta' w_i}\}$. The baseline hazard follows the Weibull distribution with $h_{u0}(t_i) = \alpha\lambda t_i^{\alpha-1}$, where $\alpha, \lambda > 0$. Then

the log-likelihood function could be written as

$$\begin{aligned}\ell(\alpha,\lambda,\boldsymbol{\gamma},\boldsymbol{\eta}|\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{X},\boldsymbol{W}) &= \sum_{i=1}^{n}\Big\{\delta_i\log\left[\pi_i f_u(t_i)\right]+(1-\delta_i)\log\left\{2E[S_M(t_i)]-1+\pi_i S_u(t_i)\right\}\Big\}\\ &= \sum_{\delta_i=1}\Big\{\log(\pi_i)+\log(\lambda)+\log(\alpha)-\lambda t_i^{\alpha}e^{\boldsymbol{\eta'w_i}}+(\alpha-1)\log t_i+\boldsymbol{\eta'w_i}\Big\}\\ &\quad+\sum_{\delta_i=0}\log\Big\{(1-\pi_i)+\pi_i\exp\left\{-\lambda t_i^{\alpha}e^{\boldsymbol{\eta'w_i}}\right\}\Big\},\end{aligned}$$

where $\pi_i = e^{\boldsymbol{\gamma'x_i}}/(0.5+e^{\boldsymbol{\gamma'x_i}})$. The maximum likelihood estimation (MLE) of $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\eta}}$, $\hat{\alpha}$ and $\hat{\lambda}$ could be obtained by implementing SAS NLMIXED procedure with respect to the above log-likelihood function. The SAS code for fitting this model is given in Appendix A.1.

For the purpose of the statistical inference, the SAS NLMIXED procedure calculate variance for the above estimations $(\hat{\alpha},\hat{\lambda},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\eta}})$ using Hessian matrix (Billingsley, 2008). The variance of parameter is $\boldsymbol{b'}\hat{\boldsymbol{H}}^{-1}\boldsymbol{b}$, where $\boldsymbol{b}$ is the resulting vector of the parameter and $\hat{\boldsymbol{H}}$ is the approximate Hessian matrix calculated in the process of maximizing the likelihood function.

3.3 Simulation

Simulation studies are conducted to evaluate the performance of proposed marginal mean survival rate model in different settings. Here we are considering two types of data generating mechanisms: (1) Data are generated from the marginal cure rate model; (2) Data are generated from the conditional cure rate model.

3.3.1 Data generated from the marginal model

In the first simulation, the survival time t_i for each individual $i(i=1,2,...,n)$ is generated from the marginal mean survival rate model with the true marginal mean survival rate,

$$E[S_M(t_i)] = \frac{1+e^{-1.5+X_{i1}+4X_{i2}}}{1+2e^{-1.5+X_{i1}+4X_{i2}}},$$

where the covariates X_{i1} and X_{i2} are independently generated from a standard normal distribution and a uniform distribution, respectively. We futher assume a Weibull as the true baseline hazard for the uncured subpopulation, thus, $h_u(t_i) = 1.1 \times 0.01 \times t_i^{0.1} \times e^{-0.5X_{i1}-2X_{i2}}$, here we consider $X = W$. Then the data generating procedure for each subject is given as follows: Each subject is grouped by a Bernoulli random variable U with the success probability π_i, and

$$\pi_i = \frac{e^{-1.5+X_{i1}+4X_{i2}}}{0.5 + e^{-1.5+X_{i1}+4X_{i2}}}$$

We know the form of π_i from Equation 3.3. If $U = 1$, the survival time of such individual is generated from the distribution associated with above $h_u(t_i)$. Otherwise, the individual is expected to have survival time $t_i = \infty$. We also generate the non-informative and independent right censoring time for each subject from an exponential distribution with three different censoring rates: (1) Heavy censoring with $\lambda_c = 0.002$; (2) Intermediate censoring with $\lambda_c = 0.001$; (3) Mild censoring with $\lambda_c = 0.0002$. The censoring rate for the uncured population is about 32.88%, 21.48% and 9.18% respectively (Appendix A.3).

Simulations are replicated 1000 times for sample sizes from 200 to 800. The results presented in Table 3.1 show that MMSR model has small bias for the parameters of interest $\boldsymbol{\gamma}$. Estimates are close to the true value of parameters and the standard errors are reduced as sample size increases. Furthermore, the coverage probability of the estimations is about desired 0.95.

Table 3.1: The parameter estimates for data generated from MMSR models with 1000 replicates

		$n = 200$			$n = 400$			$n = 600$			$n = 800$		
	True	Esti.	SE	CP	Esti.	SE	CP	Esti.	SE	CP	Esti.	SE	CP
Heavy censoring ($\lambda_c = 0.002$)													
γ_0^*	-1.5	-1.572	0.461	0.959	-1.521	0.319	0.962	-1.516	0.239	0.942	-1.511	0.210	0.956
γ_1^*	1	1.067	0.345	0.956	1.033	0.226	0.939	1.018	0.180	0.951	1.018	0.153	0.951
γ_2^*	4	4.344	1.323	0.961	4.139	0.867	0.958	4.074	0.657	0.954	4.059	0.566	0.948
η_1^*	-0.5	-0.510	0.141	0.947	-0.508	0.094	0.945	-0.507	0.078	0.952	-0.505	0.065	0.951
η_2^*	-2	-2.053	0.485	0.944	-2.015	0.328	0.945	-2.016	0.272	0.944	-2.015	0.225	0.954
α^*	1.1	1.131	0.102	0.946	1.111	0.071	0.955	1.108	0.057	0.944	1.104	0.048	0.942
λ^*	0.01	0.010	0.005	0.872	0.010	0.004	0.922	0.010	0.003	0.937	0.010	0.003	0.940
Intermediate censoring ($\lambda_c = 0.001$)													
γ_0^*	-1.5	-1.539	0.392	0.967	-1.526	0.279	0.955	-1.503	0.229	0.949	-1.510	0.194	0.951
γ_1^*	1	1.027	0.288	0.949	1.016	0.193	0.960	1.013	0.158	0.946	1.012	0.136	0.945
γ_2^*	4	4.164	1.013	0.966	4.093	0.729	0.960	4.053	0.594	0.957	4.051	0.497	0.948
η_1^*	-0.5	-0.516	0.125	0.937	-0.507	0.085	0.956	-0.508	0.070	0.940	-0.502	0.059	0.946
η_2^*	-2	-2.037	0.427	0.945	-2.021	0.289	0.959	-2.015	0.239	0.947	-2.014	0.204	0.952
α^*	1.1	1.122	0.092	0.957	1.110	0.064	0.954	1.109	0.052	0.945	1.105	0.045	0.948
λ^*	0.01	0.010	0.005	0.903	0.010	0.003	0.928	0.010	0.003	0.940	0.010	0.002	0.940
Mild censoring ($\lambda_c = 0.0002$)													
γ_0^*	-1.5	-1.525	0.398	0.949	-1.524	0.261	0.957	-1.508	0.223	0.944	-1.501	0.182	0.944
γ_1^*	1	1.040	0.272	0.943	1.015	0.174	0.945	1.011	0.146	0.953	1.012	0.124	0.949
γ_2^*	4	4.154	1.003	0.940	4.086	0.664	0.949	4.039	0.520	0.951	4.035	0.448	0.953
η_1^*	-0.5	-0.509	0.115	0.941	-0.504	0.076	0.943	-0.502	0.063	0.942	-0.501	0.054	0.949
η_2^*	-2	-2.074	0.392	0.941	-2.012	0.271	0.935	-2.015	0.215	0.942	-2.013	0.186	0.949
α^*	1.1	1.125	0.086	0.955	1.110	0.061	0.944	1.106	0.049	0.954	1.105	0.042	0.952
λ^*	0.01	0.010	0.004	0.912	0.010	0.003	0.919	0.010	0.003	0.941	0.010	0.002	0.942

CP: Coverage Probability
SE: Standard Error

3.3.2 Data generated from the conditional model

In the second simulation, we generate data from the standard conditional cure rate model, where the uncured hazard rate is $h_u(t_i) = 1.1 \times 0.01 \times t_i^{0.1} \times e^{-0.5X_{i1}-2X_{i2}}$ and the uncure fraction π_i is assumed to be a logit

$$\pi_i = \frac{e^{-1.5+Z_{i1}+4Z_{i2}}}{1+e^{-1.5+Z_{i1}+4Z_{i2}}} \tag{3.4}$$

where covariates $Z_{i1} \sim N(0,1)$ and $Z_{i2} \sim U(0,1)$. It is important to mention that $\boldsymbol{\gamma}$ is the parameter vector of interest in the marginal model, however, the true $\boldsymbol{\gamma}$ is unknown when data are generated from the conditional models. But we can derive the true $\boldsymbol{\gamma}$ by using Equation 3.3, which is

$$\frac{e^{\gamma_0+\gamma_1 Z_{i1}+\gamma_2 Z_{i1}}}{0.5+e^{\gamma_0+\gamma_1 Z_{i1}+\gamma_2 Z_{i1}}} = \frac{e^{-1.5+Z_{i1}+4Z_{i2}}}{1+e^{-1.5+Z_{i1}+4Z_{i2}}} \quad \text{holds for all } i$$

Let $Z_{i1}=0, Z_{i2}=0$, then

$$\frac{e^{\gamma_0}}{0.5+e^{\gamma_0}} = \frac{e^{-1.5}}{1+e^{-1.5}}.$$

Then $e^{\gamma_0} = 0.5e^{-1.5}$, and true $\gamma_0 = -2.193$ by solving the equation. Futhermore, we could have $\gamma_1 = 1$ and $\gamma_2 = 4$ by assuming $Z_{i1} = 0$ and $Z_{i2} = 0$, seperately. Therefore, the corresponding true parameter vector $\boldsymbol{\gamma}$ is $(-2.193,\ 1,\ 4)'$.

The censoring rate for the uncured population is about 32.88%, 21.48% and 9.18% for $\lambda_c = 0.002$, $\lambda_c = 0.001$ and $\lambda_c = 0.0002$ respectively (Appendix A.3).

Results in Table 3.2 show that MMSR model works very well to find the marginal covariate effects even the data are generated from the conditional model. The estimation bias is very small for the parameters across different settings, and the associated standard errors decrease as sample size increases.

Table 3.2: The parameter estimates for data generated from standard conditional cure rate models with 1000 replicates

		$n = 200$			$n = 400$			$n = 600$			$n = 800$		
	True	Esti.	SE	CP	Esti.	SE	CP	Esti.	SE	CP	Esti.	SE	CP
Heavy censoring ($\lambda_c = 0.002$)													
γ_0^*	-2.193	-2.281	0.474	0.964	-2.226	0.318	0.963	-2.220	0.251	0.962	-2.217	0.222	0.961
γ_1^*	1	1.071	0.341	0.954	1.028	0.219	0.951	1.025	0.173	0.952	1.026	0.149	0.953
γ_2^*	4	4.298	1.261	0.954	4.126	0.747	0.952	4.089	0.626	0.938	4.076	0.532	0.952
η_1^*	-0.5	-0.523	0.163	0.941	-0.506	0.111	0.947	-0.501	0.086	0.947	-0.506	0.076	0.948
η_2^*	-2	-2.054	0.577	0.932	-2.023	0.379	0.945	-2.036	0.305	0.951	-2.030	0.262	0.956
α^*	1.1	1.140	0.116	0.952	1.118	0.081	0.943	1.110	0.064	0.946	1.108	0.054	0.943
λ^*	0.01	0.010	0.006	0.862	0.010	0.004	0.907	0.010	0.003	0.924	0.010	0.003	0.931
Intermediate censoring ($\lambda_c = 0.001$)													
γ_0^*	-2.193	-2.253	0.438	0.953	-2.221	0.305	0.937	-2.201	0.238	0.956	-2.214	0.205	0.954
γ_1^*	1	1.054	0.292	0.948	1.023	0.192	0.951	1.022	0.152	0.953	1.015	0.132	0.956
γ_2^*	4	4.215	1.031	0.949	4.074	0.691	0.948	4.042	0.549	0.949	4.056	0.465	0.952
η_1^*	-0.5	-0.516	0.146	0.942	-0.507	0.097	0.947	-0.506	0.077	0.957	-0.502	0.067	0.939
η_2^*	-2	-2.047	0.496	0.940	-2.025	0.348	0.950	-2.007	0.267	0.951	-1.996	0.234	0.951
α^*	1.1	1.135	0.104	0.948	1.112	0.074	0.957	1.107	0.063	0.929	1.106	0.050	0.961
λ^*	0.01	0.010	0.005	0.891	0.010	0.004	0.913	0.010	0.003	0.924	0.010	0.003	0.942
Mild censoring ($\lambda_c = 0.0002$)													
γ_0^*	-2.193	-2.264	0.425	0.942	-2.205	0.272	0.957	-2.201	0.220	0.952	-2.205	0.192	0.942
γ_1^*	1	1.046	0.255	0.962	1.019	0.171	0.949	1.011	0.140	0.949	0.120	0.096	0.952
γ_2^*	4	4.188	0.955	0.943	4.046	0.588	0.959	4.028	0.497	0.950	4.031	0.417	0.950
η_1^*	-0.5	-0.519	0.129	0.941	-0.503	0.088	0.945	-0.507	0.071	0.950	-0.503	0.061	0.939
η_2^*	-2	-2.028	0.440	0.952	-2.025	0.298	0.945	-2.019	0.258	0.947	-2.014	0.212	0.944
α^*	1.1	1.123	0.095	0.956	1.109	0.068	0.956	1.110	0.053	0.952	1.106	0.047	0.953
λ^*	0.01	0.010	0.005	0.896	0.010	0.004	0.927	0.010	0.003	0.934	0.010	0.003	0.941

CP: Coverage Probability
SE: Standard Error

3.4 Discussion

This proposed MMSR method establishes a connection between covariates and the overall mean survival response. With the novel parameterization of marginal mean survival rate, we could interpret the covariate effects on the survival rate of the overall population. However, this model is naïve in the sense that the covariate effects on the marignal mean survival rate are the same as the covariate effects on the uncure fraction in the conditional cure rate model. As shown in Section 3.3.2, the only difference of above two sets of parameters or covariate effects is the intercept term, a value of $\log(0.5)$. In other words, the covariate effects on the marginal mean survival rate are equivalent to the covariate effects on the uncure fraction in the classical cure rate model.

Chapter 4

Marginal mean hazard rate model

Instead of marginal mean survival rate, we then focus on the marginal mean hazard rate. The hazard rate, or so-called failure rate, is referred to as the risk of event for an individual at a given time t. In general, it is a function of time and called hazard function. Hazard function plays a fundamental role in survival analysis as it is another representation of the distribution of survival time (Prentice and Kalbfleisch, 1979; Aalen et al., 2001). However, the marginal hazard function for the cure rate model is still not standard. Therefore, we consider to use the marginal mean hazard rate which could be obtained by taking the expectation of the marginal hazard function with respect to time t, which is similar to mean hazard rate discussed by Makino (1984). This marginal mean hazard rate then represents the risk of event on average. We can relate it to covariates via a proper link function. We call this reparameterized cure rate model as marginal mean hazard rate model (MMHR), which provides a straightforward interpretation of covariate effects on the overall population, specifically, on the marginal mean hazard rate.

4.1 Formulation of marignal mean hazard rate model

In this study, we propose a marginal mean hazard rate model which establish a direct connection between the marginal mean hazard rate $E[h_M(t_i)]$ and the covariates through a

link function. A reasonable link function should have the same range as the marginal mean hazard rate, i.e. $E[h_M(t_i)] > 0$. We assume a logit link function for marginal mean hazard rate,

$$E[h_M(t_i)] = e^{\boldsymbol{\beta' x_i}} \tag{4.1}$$

The parameter of interest is the $\boldsymbol{\beta}$ which can be used to interpret the covariate effects directly on the marginal mean hazard rate. To estimate $\boldsymbol{\beta}$, we use the framework of the standard cure rate model with the assumption of Weibull baseline hazard. Specifically, we would replace the conditional hazard rate with the function of the marginal mean hazard rate.

Let U_i be the unobserved indicator, where $U_i = 1$ if the individual i is uncured with probability π_i and $U_i = 0$ if cured with probability $1 - \pi_i$. Then the marginal hazard rate could be written as follows, we have

$$h_M(t_i) = \begin{cases} h_u(t_i), & \text{if} \quad U_i = 1 \\ 0, & \text{if} \quad U_i = 0 \end{cases}$$

Considering t_i as a continuous random variable and then taking the expectation on marginal hazard rate with respect to t,

$$\begin{aligned} E[h_M(t_i)] &= E_{t_i}\Big\{E_{U_i}\Big[h_M(t_i)|U_i\Big]\Big\} \\ &= E_{t_i}[\pi_i h_u(t_i)] \\ &= \pi_i E[h_u(t_i)] \end{aligned} \tag{4.2}$$

As we assumed in a standard cure rate model, the uncure fraction is $\pi_i = e^{\boldsymbol{\xi' z_i}}/(1 + e^{\boldsymbol{\xi' z_i}})$ and the conditional hazard function $h_u(t_i) = \alpha\lambda t_i^{\alpha-1} e^{\boldsymbol{\eta' w_i}}$. Since we are not interested in $\boldsymbol{\eta}$,

we can rewrite $\boldsymbol{\eta'w_i}$ as u_i. Then by Equations 4.1 and 4.2,

$$\begin{aligned} e^{\boldsymbol{\beta' x_i}} &= \pi_i E[h_u(t_i)] \\ &= \pi_i \int \alpha\lambda t_i^{\alpha-1} e^{\mu_i} f_u(t_i) dt_i \\ &= \pi_i \alpha\lambda e^{\mu_i} \int t_i^{\alpha-1} \lambda\alpha e^{\mu_i} t_i^{\alpha-1} \exp(-\lambda t_i^{\alpha} e^{\mu_i}) dt_i \\ &= \pi_i \alpha\lambda e^{\mu_i} E[t_i^{\alpha-1}] \\ &= \pi_i \alpha [\lambda e^{\mu_i}]^{\frac{1}{\alpha}} \Gamma(2 - \frac{1}{\alpha}) \end{aligned}$$

The last equality comes from the fact that $E[t_i^k] = (\lambda e^{\mu_i})^{-\frac{k}{\alpha}} \Gamma(1 + \frac{k}{\alpha})$, the $k-$th moment functions of Weibull distribution. The previous equation can be rewritten as

$$\lambda e^{\mu_i} = \Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha \pi_i \Gamma(2 - \frac{1}{\alpha})} \Big]^{\alpha}, \tag{4.3}$$

and we then can replace λe^{μ_i} in the likelihood function of standard cure rate model by the right side of Equation 4.3. It is worth to mention that since we are not interested in $\boldsymbol{\eta}$, and therefore μ_i could be a more general form rather than a linear function of covariates $\boldsymbol{w_i}$.

4.2 Likelihood function and estimation

Suppose we observe independent data with $\{t_i, \delta_i, \boldsymbol{x_i}, \boldsymbol{z_i}\}$ for i^{th} subject, where these notations are the same as we defined in Chapter 3. The likelihood function for marginal mean

hazard rate model is given by

$$\begin{aligned}
\mathcal{L}(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi} | t_1, ..., t_n) &= \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left[\pi_i f_u(t_i)\right]^{\delta_i} \left[1 - \pi_i + \pi_i S_u(t_i)\right]^{1-\delta_i} \\
&= \prod_{\imath=1}^{n} \left\{ \pi_i \alpha \Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha \pi_i \Gamma(2 - \frac{1}{\alpha})} \Big]^{\alpha} t^{\alpha - 1} \exp\left\{ - t^{\alpha} \Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha \pi_i \Gamma(2 - \frac{1}{\alpha})} \Big]^{\alpha} \right\} \right\}^{\delta_i} \\
&\quad \left\{ 1 - \pi_i + \pi_i \exp\left\{ - t^{\alpha} \Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha \pi_i \Gamma(2 - \frac{1}{\alpha})} \Big]^{\alpha} \right\} \right\}^{1-\delta_i}
\end{aligned}$$

where $\pi_i = e^{\boldsymbol{\xi' z_i}}/(1 + e^{\boldsymbol{\xi' z_i}})$.

We can estimate the above parameters through the regular likelihood approach and the SAS code for fitting this model is given in Appendix A.2. Moreover, we could obtain the standard error of above estimates by Hessian matrix for the purpose of statistical inference. The equation for the Hessian matrix could be referred to Section 3.2.

4.3 Simulation

Here we are considering two types of data generating mechanisms: (1) Data are generated from the marginal mean hazar rate model; (2) Data are generated from the standard conditional cure rate model.

4.3.1 Data generated from the marginal model

In the first simulation, the survival time $t_i (i = 1, 2, ..., n)$ is generated from the marginal cure rate model with the true $E[h_M(t_i)] = \exp\{-4 + X_{i1} - 2X_{i2}\}$ and $X_{i1} \sim N(0, 1)$ and

$X_{i2} \sim U(0,1)$. By Equation 4.3, the hazard function of uncured subpopulation is given as

$$\begin{aligned} h_u(t_i) &= \alpha t_i^{\alpha-1}\Big[\frac{e^{\boldsymbol{\beta}'\boldsymbol{x_i}}}{\alpha\pi_i\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha} \\ &= 0.75 \times t_i^{-0.25}\Big[\frac{\exp\{-4+X_{i1}-2X_{i2}\}}{0.75\times\pi_i\times\Gamma(2/3)}\Big]^{0.75}, \end{aligned}$$

where $\text{logit}(\pi_i) = 1 + 2X_{i1} + 1.5X_{i2}$. Based on the above setting, the data generating procedure is the same as we described in Section 3.3.1. The censoring time are generated from an exponential distribution with rate λ_c taking values on 0.0002, 0.001, 0.002, representing mild, intermediate and heavy censoring. The true censoring rate for the uncured population is about 18.06%, 13.01% and 7.67% respectively (Appendix A.3). Simulations are replicated 1000 times for sample size from 200 to 800.

The results are presented in Table 4.1, which show the average estimations for all parameters are close to the true value. We could also see the accuracy of estimation increases and standard error decreases as the sample size gets larger. Moreover, coverage probability is very close to 0.95 as expected. Those findings indicate that our maximum likelihood estimators should be asymptotically consistent.

Table 4.1: Marginalized mean hazard rate model with 1000 simulations and varying sample size and censoring rate (Data generated from the marginalized model

		$n = 200$			$n = 400$			$n = 600$			$n = 800$		
	Value	Esti.	SE	CP	Esti.	SE	CP	Esti.	SE	CP	Esti.	SE	CP
Heavy censoring ($\lambda_c = 0.002$)													
α^*	0.75	0.764	0.055	0.957	0.757	0.038	0.945	0.754	0.034	0.955	0.752	0.027	0.943
β_0^*	-4	-3.985	0.222	0.947	-3.993	0.156	0.945	-3.995	0.127	0.957	-3.967	0.110	0.950
β_1^*	1	0.999	0.264	0.940	0.996	0.187	0.951	1.000	0.152	0.957	1.002	0.132	0.950
β_2^*	-2	-2.018	0.158	0.931	-2.009	0.111	0.945	-2.000	0.091	0.961	-2.007	0.079	0.953
ξ_0^*	1	1.064	0.349	0.953	1.016	0.238	0.961	1.022	0.194	0.951	1.012	0.167	0.949
ξ_1^*	2	2.113	0.518	0.959	2.043	0.353	0.945	2.030	0.286	0.947	2.018	0.246	0.942
ξ_2^*	1.5	1.603	0.365	0.962	1.539	0.245	0.957	1.534	0.199	0.956	1.519	0.170	0.948
Intermediate censoring ($\lambda_c = 0.001$)													
α^*	0.75	0.768	0.053	0.951	0.757	0.037	0.945	0.756	0.030	0.946	0.753	0.026	0.932
β_0^*	-4	-3.981	0.211	0.945	-3.999	0.151	0.951	-4.001	0.123	0.942	-3.986	0.107	0.944
β_1^*	1	0.983	0.251	0.943	1.003	0.179	0.952	1.007	0.146	0.950	1.003	0.126	0.950
β_2^*	-2	-2.008	0.149	0.948	-2.003	0.106	0.947	-2.001	0.086	0.959	-2.004	0.075	0.938
ξ_0^*	1	1.062	0.321	0.954	1.024	0.222	0.960	1.022	0.180	0.962	1.018	0.155	0.937
ξ_1^*	2	2.073	0.503	0.957	2.060	0.346	0.955	2.025	0.279	0.954	2.016	0.241	0.969
ξ_2^*	1.5	1.584	0.337	0.945	1.552	0.232	0.939	1.530	0.186	0.955	1.523	0.160	0.950
Mild censoring ($\lambda_c = 0.0002$)													
α^*	0.75	0.763	0.051	0.948	0.755	0.036	0.950	0.755	0.032	0.949	0.754	0.025	0.939
β_0^*	-4	-3.987	0.207	0.950	-3.991	0.147	0.951	-3.995	0.120	0.955	-4.002	0.103	0.952
β_1^*	1	1.001	0.242	0.957	1.003	0.172	0.943	0.995	0.140	0.950	1.007	0.121	0.949
β_2^*	-2	-2.007	0.143	0.958	-2.007	0.101	0.944	-2.003	0.082	0.965	-2.004	0.071	0.957
ξ_0^*	1	1.052	0.299	0.964	1.016	0.206	0.951	1.003	0.167	0.952	1.008	0.145	0.953
ξ_1^*	2	2.087	0.492	0.965	2.039	0.337	0.939	2.042	0.274	0.950	2.018	0.235	0.955
ξ_2^*	1.5	1.579	0.318	0.960	1.535	0.216	0.952	1.524	0.175	0.958	1.519	0.150	0.955

CP: Coverage Probability
SE: Standard Error

4.3.2 Data generated from the conditional model

This simulation is conducted to evaluate the performance of proposed marginal mean hazard rate method with the Average Predicted value (APV) approach proposed by Albert et al. (2014). The survive time t_i is generated from the conditional cure rate model with

$$\begin{aligned} h_u(t_i) &= h_{u0}(t_i) \exp\left\{\boldsymbol{\eta' x_i}\right\} \\ &= 0.75 \times 0.1 \times t_i^{-0.25} \times \exp\{-4 + X_{i1} - 2X_{i2}\} \end{aligned}$$

where X_{i1} is an binary exposure covariate taking value 1 if subject is exposed and 0 otherwise. For X_{i2}, we consider to use a standard normal distribution or a uniform distribution. The uncure fraction is assumed to be $\text{logit}(\pi_i) = 1 + 2X_{i1} + 1.5X_{i2}$. Censoring times is generated from a exponential distribution with $\lambda_c = 0.0002$. As we are assuming the marginal mean hazard rate $E[h_M(t)] = e^{\boldsymbol{\beta' x}}$ in our model, the true parameter $\boldsymbol{\beta}$ is actually unknown when data generated from a conditional cure rate model. Instead, we evaluate the ratio of two marginal mean hazard rates for exposure and nonexposure group after averaging out the extraneous covariates as suggested by the APV method. Specifically, the true ratio can be obtained by

$$\begin{aligned} \theta_R &= \frac{\int E[h_M(t)|x_1 = 1, x_2]dx_2}{\int E[h_M(t)|x_1 = 0, x_2]dx_2} \\ &= \frac{\int \frac{e^{\xi_0+\xi_1+\xi_2 x_2}}{1+e^{\xi_0+\xi_1+\xi_2 x_2}} \left(e^{\eta_1+\eta_2 x_2}\right)^{\frac{1}{\alpha}} f(x_2)dx_2}{\int \frac{e^{\xi_0+\xi_2 x_2}}{1+e^{\xi_0+\xi_2 x_2}} \left(e^{\eta_2 x_2}\right)^{\frac{1}{\alpha}} f(x_2)dx_2}, \end{aligned} \tag{4.4}$$

where $f(x_2)$ is the pdf function of the covariate X_2. The last equality holds in Equation 4.4 since

$$\begin{aligned} &\int E[h_M(t)|x_1 = 1, x_2]dx_2 \\ &= \int \pi_i \alpha [\lambda e^{\mu_i}]^{\frac{1}{\alpha}} \Gamma(2 - \frac{1}{\alpha}) f(x_2)dx_2 \\ &= \int \frac{e^{\xi_0+\xi_1+\xi_2 x_2}}{1 + e^{\xi_0+\xi_1+\xi_2 x_2}} \alpha \left[\lambda e^{\eta_0+\eta_1+\eta_2 x_2}\right]^{\frac{1}{\alpha}} \Gamma\left(2 - \frac{1}{\alpha}\right) f(x_2)dx_2 \end{aligned}$$

and

$$\begin{aligned}
&\int E[h_M(t)|x_1 = 0, x_2]dx_2 \\
&= \int \pi_i \alpha [\lambda e^{\mu_i}]^{\frac{1}{\alpha}} \Gamma(2 - \frac{1}{\alpha}) f(x_2) dx_2 \\
&= \int \frac{e^{\xi_0 + \xi_2 x_2}}{1 + e^{\xi_0 + \xi_2 x_2}} \alpha \Big[\lambda e^{\eta_0 + \eta_2 x_2}\Big]^{\frac{1}{\alpha}} \Gamma\Big(2 - \frac{1}{\alpha}\Big) f(x_2) dx_2
\end{aligned}$$

On the other hand, we could get the same mean ratio easily through the proposed marginal mean hazard rate that is defined in Equation 4.1. Specifically,

$$\begin{aligned}
\theta_{RM} &= \frac{E[h_M(t)|x_1 = 1, x_2]}{E[h_M(t)|x_1 = 0, x_2]} \\
&= \frac{\exp\{\beta_0 + \beta_1 + \beta_2 x_2\}}{\exp\{\beta_0 + \beta_2 x_2\}} \\
&= \exp\{\beta_1\}
\end{aligned}$$

Therefore, we can estimate the β_1 through the proposed model to further obtain $\hat{\theta}_{RM}$. Then the $\hat{\theta}_{RM}$ can be evaluated with the true θ_R obtained by the APV method.

Simulations are replicated 1000 times for sample size from 200 to 800. We implement the simulations by using the Quasi-Newton nonlinear optimization embedded in SAS 9.4 NLMIXED procedure (SAS Institue, Cary, NC, USA).

The simulation results presented in Table 4.2 indicate that proposed method perform well in estimating the overall exposure effect on marginal mean hazard. As sample size increases, the estimated mean ratio is getting close to the true mean ratio and corresponding standard errors decreases.

4.4 Application to SEER liver cancer data

We apply the marginal mean hazard rate model to the liver cancer data collected from the state of Connecticut by the Surveillance, Epidemiology, and End Results (SEER) Program, which is the most comprehensive source of population-based cancer data in the United States.

Table 4.2: Simulation results for mean ratio (exposed vs nonexposed) when data generated from the conditional model with two covariates and $\lambda_c = 0.0002$

	$X_1 \sim Binom(n, 0.5)$			
	$X_2 \sim N(0,1)$		$X_2 \sim U(0,1)$	
n	AVE EST	SE	AVE EST	SE
200	4.884	0.635	4.695	0.473
400	4.730	0.590	4.740	0.460
600	4.778	0.524	4.741	0.449
800	4.850	0.442	4.772	0.421
True mean ratio θ_R	4.845		4.781	

In our dataset, we have total 2362 patients age from 10 to 96 who were diagnosed with liver cancer between 1975 and 2016. The event in this study is the death of liver cancer. Our primary objective in this study is to evaluate the covariate effects such as age or surgery treament on the overall mean hazard.

As seen from Figure 4.1, survival curve levels off at the rate about 0.4 after 180 months of follow-up, which shows an evidence of the existence of long-term survivors. Furthermore, we conduct a non-parametric hypothesis test which is developed by Maller and Zhou (1994) whether the follow-up is sufficiently long for cure rate model. In this test, the p-value is $\hat{p}_n = (1 - N_n/n)^n$, where N_n is the number of observed events between the interval $(2T_n^* - T_n, T_n^*)$, and T_n^* denote the largest uncensored survival time in the dataset and T_n represent the largest censored survival time in the dataset. Here $n = 2362$, $N_n = 14$, where N_n is the number of observed events between the interval $(2T_n^* - T_n, T_n^*) = (171, 247)$. $T_n^* = 247$ is the largest uncensored survival time in the dataset and $T_n = 323$ is the largest censored survival time in the dataset. Thus $\hat{a} = 7.98 \times 10^{-7} < 0.05$, leads to the conclusion that follow-up time is long enough to observe the cured patients.

As long-term survivors are justified, we apply the proposed methods to the data with covariates of patient's Age at diagnosis (Age), Surgery or not (SUR) and Number of lymph nodes (NLN). Table 4.3 summarizes the information of those covariates. The marginal mean

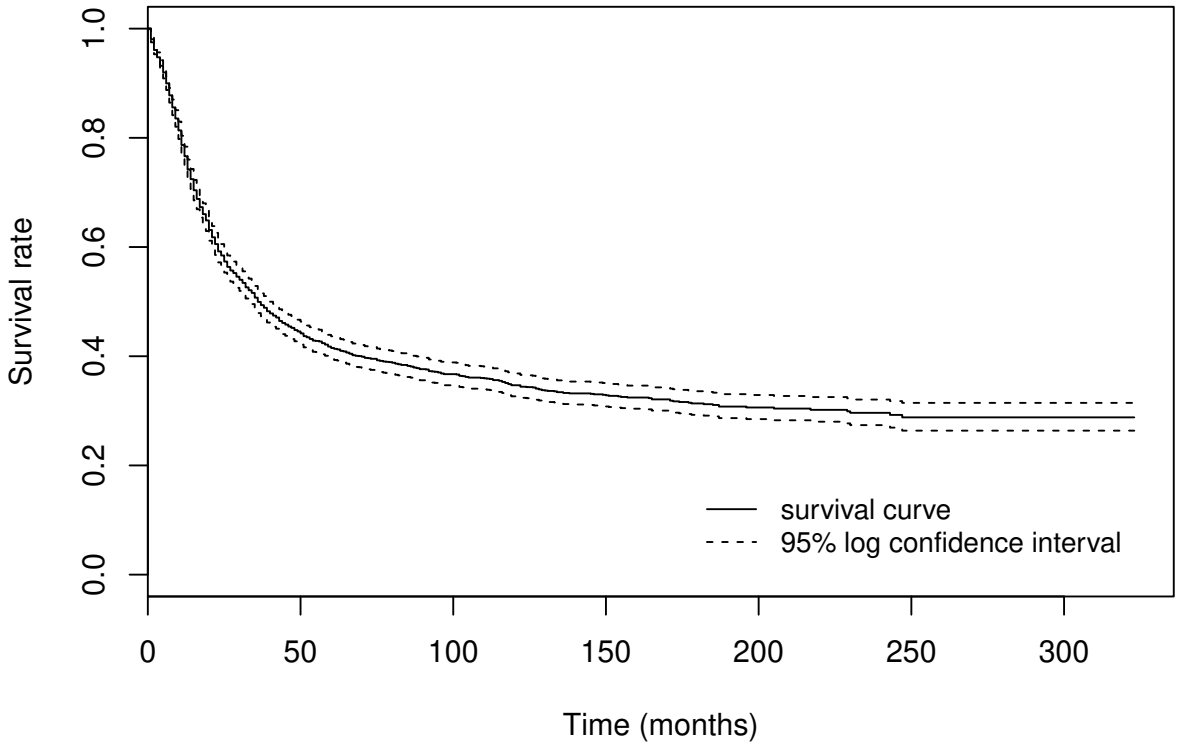


Figure 4.1: Survival curve and 95% log confidence interval for the liver cancer patients in SEER registry

Table 4.3: Summary statistics of covariates for 2362 liver patients in SEER registry

Name		Mean	Std Dev	Min.	Max.
Age at diagnosis (years)		66.47	12.52	10	96
# of lymph nodes		1.43	1.90	0	6
Surgery			Frequency		Percent(%)
	Yes		2278		96.44
	No		84		3.56
	Total		2362		100

hazard rate and uncure fraction are related to covariates as follows,

$$E[h_M(t_i)] = \exp\left\{\beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{SUR}_i + \beta_3 \text{NLN}_i\right\} \quad \text{and}$$
$$\text{logit}(\pi_i) = \xi_0 + \xi_1 \text{AGE}_i + \xi_2 \text{SUR}_i + \xi_3 \text{NLN}_i$$

The estimates and the standard errors are given in Table 5.5. As seen, covariates Age and Number of lymph nodes have positive effects on the overall mean hazard rate ($\hat{\beta}_1 = 0.117$, p-value < 0.001; $\hat{\beta}_3 = 0.249$, p-value < 0.001), which means that the patients diagnosed at older age with higher number of lymph nodes might expose to a higher risk of death. Surgery, however, is negatively associated with the marginal mean hazard rate ($\hat{\beta}_2 = -0.948$, p-value < 0.001), meaning that the liver cancer patients who received the surgery would expect to have a lower risk of death. This application is an excellent example where the classical formulation of cure rate model fails to explain the overall effect of covariates on the overall mean hazard rate directly.

4.5 Discussion

To our best knowledge, this is the first study proposing a marginal method to interpret the covariate effects on the overall hazard for the two-component mixture cure rate model. This method is particularly useful when the aim is the statistical inference on the overall mean

Table 4.4: The parameter estimations for the liver cancer patients using the marginalized mean hazard rate model with Weibull baseline hazard assumption

	Covariate	Parameter	Estimate	SE	p-value
Marginal mean hazard	Age	β_1	0.117	0.035	< 0.001
	Surgery	β_2	-0.948	0.135	< 0.001
	Lymph Nodes	β_3	0.249	0.029	< 0.001
Noncure fraction	Age	ξ_1	0.201	0.052	< 0.001
	Surgery	ξ_2	-1.306	0.373	< 0.001
	Lymph Nodes	ξ_3	0.395	0.058	< 0.001
Weibull baseline hazard	Shape	α	0.982	0.021	< 0.001

hazard rate. However, the assumption of the conditional baseline hazard in the model is too restrictive as we only assume the Weibull. It is challenging to relax this assumption by using more general distributions or nonparametric techniques due to potential difficulties in the calculation of intergral. This will be a topic for the next chapter. Another chanllenging but interesting extension is how to apply the proposed marginal model to high-dimensional data with a proper variable selection procedure. We will discuss about this study in Chapter 6.

Chapter 5

Semi-parametric marginal mean hazard model

The proposed marginal cure rate model in Chapter 4 assumes the uncured subpopulation has a parametric Weibull baseline hazard function for its generality as well as simplicity. However, this assumption could be too restrictive in practice, particularly as the baseline hazard is often unknown *a priori*. To relax this assumption, a semi-parametric method is proposed for the marginal mean hazard models, where the baseline hazard is nonparametric and the marginal mean hazard is denoted by a parametric link function.

The semi-parametric methods for the regular cure rate models have been well discussed in the literature. For example, Kuk and Chen (1992) proposed a semi-parametric approach in which a Cox proportional hazard model was used to incorporate the covariate information but its baseline hazard was estimated nonparametrically. Taylor (1995) used an EM algorithm and a Kaplan-Meier type approach to estimate the uncured survival curve. Similarly, Sy and Taylor (2000), and Peng and Dear (2000) proposed semi-parametric models where the cumulative hazards for the uncured subpopulation can be estimated nonparametrically by using Breslow or Aalon-Nelson estimator. Liu and Shen (2009) presented a semi-parametric cure model for the regression analysis of interval-censored event data, and the semi-parametric maximum likelihood estimation was obtained by using the EM method.

Wang et al. (2012) further applied spline approaches to model both the cure fraction and conditional survival function of uncured subpopulation. Zhou et al. (2017) presented a class of semi-parametric transformation models for the interval-censored survival data and developed a sieve maximum likelihood approach to estimate model.

The key characteristic of the aforementioned studies for cure rate model is that EM algorithm is naturally used such that the regression parameters for the cure fraction and the baseline hazard for the uncured population could be estimated separately in the M step. Then the classical nonparametric method, such as Kaplan-Meier curve, Breslow or Aalon-Nelson estimator, could be directly applied to model the uncured subpopulation. However, to our best knowledge, there is no study in the literature to discuss the semi-parametric or nonparametric marginal cure rate models. Moreover, according to our study, EM algorithm is not helpful for the marginal cure rate model as it cannot separate the marginal parameters and the distribution of uncured subpopulation in the M step. Thus we could not enjoy the benefits of fitting each component separately by using EM algorithm. Appendix A.4 provides detailed evidence about this statement. Our objective is to propose a new semi-parametric method for the marginal cure rate model, which could model the uncure baseline hazard nonparametrically and keep the marginal mean hazard rate being parametric for the interpretation.

In Chapter 4, we assume a Weibull function as the baseline hazard function of the uncured subpopulation $h_{u0}(t)$. As a result, the analytic form of $E[h_{u0}(t)]$ could be derived with some efforts. With all the components in the likelihood function being parameteric, the regular maximum likelihood method could be employed for the model estimation. As we mentioned previously, however, this assumption could be too restrictive in practice. To relax this restriction, we propose a new semi-parametric approach for our marginal cure rate mode. Specifically, Bernstein polynomials can be used to model the uncured cumulative hazard function $H_{u0}(t)$.

5.1 Bernstein polynomials

In this section, we introduce a semi-parametric method to model the uncured hazard function of marginal cure rate models by using Bernstein polynomials. A Bernstein polynomial is a technique using the linear combination of Bernstein basis to approximate functions. Carnicer and Pena (1993) gave an affirmative statement that the Bernstein polynomials has the optimal shape preserving property among all other polynomial approximation. Delgado and Pena (2012) further demonstrated the optimal stability property of Bernstein polynomials for the fastest convergence rates of the corresponding iteration approximation. Osman and Ghosh (2012) pointed out that Bernstein polynomials could naturally capture the monotonically non-decreasing property of the cumulative hazard function, and have nice differentiability properties such that the log-likelihood takes easy forms, making the implementation is relatively efficient as compared to other computationally intensive splines. Owing to its good properties, Zhou et al. (2017) employed the Bernstein polynomials to model interval-censored failure time data, which could provide additional flexibility to classical proportional hazard models.

5.2 Marignal mean hazard model with Bernstein polynomials

The log likelihood function $\ell(\boldsymbol{\beta}, \boldsymbol{\xi}, H_{u0}(t))$ of our marginal cure rate model consists of marginal parameter vector $\boldsymbol{\beta}$, cure rate parameter vector $\boldsymbol{\xi}$ and unknown cumulative hazard function $H_{u0}(t)$. Since we employ a nonparametric approach for the cumulative hazard function $H_{u0}(t)$, the likelihood function $\ell(\boldsymbol{\beta}, \boldsymbol{\xi}, H_{u0}(t))$ will involve not only the finite-dimensional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, but also the infinite-dimensional $H_{u0}(t)$. To alleviate the dimensional issue of $H_{u0}(t)$, we refer to Huang and Rossini (1997); Osman and Ghosh (2012); Zhou et al. (2017) and consider a sieve likelihood method for the estimation based on Bernstein polynomials. According to Zhou et al. (2017), the sieve space M for the infinite-dimensional $H_{u0}(t)$

is defined as

$$M = \left\{ H_{u0}(t) = \sum_{k=0}^{m} \phi_k B_k(x, m),\ 0 \leq \phi_1 \leq \phi_2 \leq \cdots \leq \phi_m \right\} \tag{5.1}$$

where M represents the reduced sieve space for $H_{u0}(t)$, and $B_k(x, m)$ are basis of Bernstein polynomials with degree $m = o(n^{\nu})$ for some $\nu \in (0, 1)$. Specifically,

$$B_k(x, m) = \binom{m}{k} (x)^k (1-x)^{m-k}, k = 0, 1, 2..., m, \text{ and } x = \frac{t}{\max(t)} \tag{5.2}$$

Note that the cumulative hazard function in Equation 5.1 is positive and non-decreasing. Thus, some constraints could be imposed on the coefficients of Bernstein polynomials, such as $0 \leq \phi_0 \leq \phi_1 ... \leq \phi_m$. Furthermore, the basis for baseline hazard function $h_{u0}(t)$ could be obtained by taking the first derivative of Equation 5.2 with respect to x. The form is given as,

$$b_k(x, m) = \binom{m}{k} k x^{k-1} (1-x)^{m-k} - \binom{m}{k} (m-k) x^k (1-x)^{m-k-1}, \quad k = 0, 1, 2, ..., m,$$

where $\binom{m}{k+1}(k+1) = \binom{m}{k}(m-k) = \frac{m!}{k!(m-k-1)!}$. Then by the inductive method, we can obtain the baseline hazard

$$h_{u0}(x) = \sum_{k=1}^{m} (\phi_k - \phi_{k-1}) \frac{m!}{(k-1)!(m-k)!} \left[x^{k-1} (1-x)^{m-k} \right].$$

As an example, the Bernstein approximation with degree of 5 for $H_{u0}(x)$ is,

$$\begin{aligned} H_{u0}(x) &= B(x, 5) \\ &= \sum_{k=0}^{5} \phi_k B_k(x, 5) \\ &= \phi_0 (1-x)^5 + 5\phi_1 x (1-x)^4 + 10\phi_2 x^2 (1-x)^3 + 10\phi_3 x^3 (1-x)^2 + 5\phi_4 x^4 (1-x) + \phi_5 x^5 \end{aligned}$$

and its associated baseline hazard function $h_{u0}(x)$ is

$$\begin{aligned} h_{u0}(x) &= -5\phi_0(1-x)^4 + 5\phi_1(1-x)^4 - 20\phi_1 x(1-x)^3 + 20\phi_2 x(1-x)^3 - 30\phi_2 x^2(1-x)^2 \\ &\quad + 30\phi_3 x^2(1-x)^2 - 20\phi_3 x^3(1-x) + 20\phi_4 x^3(1-x) - 5\phi_4 x^4 + 5\phi_5 x^4 \\ &= 5(\phi_1-\phi_0)(1-x)^4 + 20(\phi_2-\phi_1)x(1-x)^3 + 30(\phi_3-\phi_2)x^2(1-x)^2 \\ &\quad + 20(\phi_4-\phi_3)x^3(1-x) + 5(\phi_4-\phi_3)x^4 \end{aligned}$$

Then we could get the sieve likelihood function of the marginal mean hazard rate model by substituting $h_{u0}(t)$ and $H_{u0}(t)$ with these Bernstein approximations. It is necessary to determine the degree of Bernstein polynomials in practice. Zhou et al. (2017) suggested to choose the degree m which could minimize

$$\text{AAIC} = -2\ell(\hat{\theta}) + 2(p + 1 + 2(m+1))$$

where AAIC is adjusted Akaike information criterion used for model selection (Anderson and Burnham, 2004), $\hat{\theta}$ is the estimation of parameters and p is the number of parameters used in the model. It is worth mentioning that using a higher value of m can improve the approximation but it would be less efficient in the computation.

5.3 Likelihood function and estimation

After employing the Bernstein approximation for the baseline hazard function, $E[h_{u0}(t)]$ appeared in the likelihood equation needs to be handled. This expectation is hard to derive if the density of x, or $t/\max(t)$ has a complicated form. We note that Bernstein approximation of baseline hazard function is

$$h_{u0}(x) = \sum_{k=1}^{m} (\phi_k - \phi_{k-1}) \frac{m!}{(k-1)!(m-k)!} x^{k-1}(1-x)^{m-k}, \quad x = \frac{t}{\max(t)}$$

which has the closed form for any fixed m and has the kernel of Beta distribution. To calculate the $E[h_{u0}(t)]$, we then assume a Beta distribution for $f(x)$, which has the following advantages. First, it could help deriving the closed form of $E[h_{u0}(t)]$ as shown in Equation 5.3. Second, this assumption brings two more free parameters a and b to the $E[h_{u0}(t)]$, which should be flexible enough to account for $E[h_{u0}(t)]$. Lastly, this assumption is only used for calculating $E[h_{u0}(t)]$. The likelihood function is still derived based on the Bernstein polynomials according to Equation 5.4. As

$$f(x) = \frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)}, \quad a > 0,\ b > 0,\ 0 < x < 1$$

then we could derive the closed form of $E[h_{u0}(t)]$. For any fixed m, $E[h_{u0}(t)]$ could be expressed as

$$\begin{aligned}
E[h_{u0}(t)] &= \int_0^1 \sum_{k=1}^{m} (\phi_k - \phi_{k-1}) \frac{m!}{(k-1)!(m-k)!} x^{k-1}(1-x)^{m-k} f(x)dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \sum_{k=1}^{m} \frac{(\phi_k - \phi_{k-1})m!}{(k-1)!(m-k)!} x^{k-1}(1-x)^{m-k} x^{a-1}(1-x)^{b-1}dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \sum_{k=1}^{m} \frac{(\phi_k - \phi_{k-1})m!}{(k-1)!(m-k)!} x^{k+a-2}(1-x)^{m+b-k-1}dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \sum_{k=1}^{m} \frac{(\phi_k - \phi_{k-1})m!}{(k-1)!(m-k)!} \frac{\Gamma(k+a-1)\Gamma(m+b-k)}{\Gamma(m+a+b-1)} \\
&= \frac{\sum_{k=1}^{m} \frac{(\phi_k - \phi_{k-1})m!}{(k-1)!(m-k)!} \Gamma(k+a-1)\Gamma(m+b-k)(a+b-1)!}{\Gamma(a)\Gamma(b)(m+a+b-2)!}
\end{aligned} \tag{5.3}$$

Then, suppose we observe independent data with $\{t_i, \delta_i, \boldsymbol{x_i}, \boldsymbol{z_i}\}$ for i^{th} subject, the likelihood function of the proposed semi-parametric marginal mean hazard rate model is

$$\begin{aligned}
\mathcal{L}(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi} | t_1, ..., t_n) = \prod_{i=1}^{n} &\left\{ h_{u0}(t_i) \frac{e^{\boldsymbol{\beta' x_i}}}{E[h_{u0}(t)]} \exp\left\{ - H_{u0}(t_i) \frac{e^{\boldsymbol{\beta' x_i}}}{\pi_i E[h_{u0}(t)]} \right\} \right\}^{\delta_i} \times \\
&\left\{ 1 - \pi_i + \pi_i \exp\left\{ - H_{u0}(t_i) \frac{e^{\boldsymbol{\beta' x_i}}}{\pi_i E[h_{u0}(t)]} \right\} \right\}^{1-\delta_i}
\end{aligned} \tag{5.4}$$

where

$$\pi_i = \frac{e^{\boldsymbol{\xi' z_i}}}{1 + e^{\boldsymbol{\xi' z_i}}}$$

and

$$H_{u0}(t_i) = \sum_{k=0}^{m} \phi_k \binom{m}{k} (x_i)^k (1 - x_i)^{m-k}, k = 0, 1, 2..., m, \text{ and } x_i = \frac{t_i}{\max(t_i)}$$

and

$$E[h_{u0}(t)] = \frac{\sum_{k=1}^{m} \frac{(\phi_k - \phi_{k-1})m!}{(k-1)!(m-k)!} \Gamma(k + a - 1)\Gamma(m + b - k)(a + b - 1)!}{\Gamma(a)\Gamma(b)(m + a + b - 2)!} \tag{5.5}$$

With above Equation 5.4 and 5.5, we can estimate the parameters of interest $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ through the likelihood approach. Zhou et al. (2017) proved that the estimators of sieve likelihood estimation using Bernstein ploynomials enjoy the property of asymptotic consistency and normality when $n \to \infty$. According to this paper, assuming the regularity conditions held, and the cumulative baseline hazard $H_{u0}(t)$ is continuously differentiable up to order γ, $\gamma \geq 2$, then for estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\xi}}_n$, we have

$$\sqrt{n}\{(\hat{\boldsymbol{\beta}}_n', \hat{\boldsymbol{\xi}}_n')' - (\boldsymbol{\beta}_0', \boldsymbol{\xi}_0')'\} \to_d N\{0, I^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\}$$

where $I^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)$ is bounded. Namely, for any vection $\boldsymbol{b}$ with $\|\boldsymbol{b}\| \leq 1$, there exists $v^* \in \bar{V}$ such that $\|v^*\| = \boldsymbol{b}' I^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\boldsymbol{b}$. $\bar{V}$ is the closed linear span of parameter space. Moreover, the degree of Bernstein polynomial m will determine the rate of convergence as $d\big[(\hat{\boldsymbol{\beta}}_n', \hat{\boldsymbol{\xi}}_n'), (\boldsymbol{\beta}_0', \boldsymbol{\xi}_0')\big] = O_p(n^{-min\{(1-\nu)/2, \nu r/2\}})$, where $\nu \in (0, 1)$ such that $m = o(n^\nu)$. The regularity conditions are detailed in the appendix of Zhou et al. (2017). The parameter estimators could be further evaluated by statistical inference after getting the standard error from the Hessian matrix. In this study, we will implement the maximization of above likelihood function using *maxLik* package of *R* software.

5.4 Identifiability problem

It should be noted that the identifiability problem is present for the intercept term β_0 after introducing the Bernstein polynomials for the baseline hazard function. The Profile Likelihood (PL) approach described by Raue et al. (2014) can be used to detect the non-identifiable parameters. The idea of this approach is that changing the value of non-identifiable parameter doesn't have an impact on the maximized likelihood function or likelihood profile. This profile could be obtained for each parameter, say θ_j individually by repeating the maximization for a series of different value of the parameter θ_j, namely,

$$PL(\theta_i) = \max_{j \neq i}[L(\theta_j)]$$

As an example, Figure 5.1 shows likelihood profiles of marginal mean hazard parameters and uncure fraction parameters. A flat profile demonstrates a structurally non-identifiable parameter. It shows that only intercept term β_0 is structurally not identifiable. (Witten and Tibshirani, 2010)

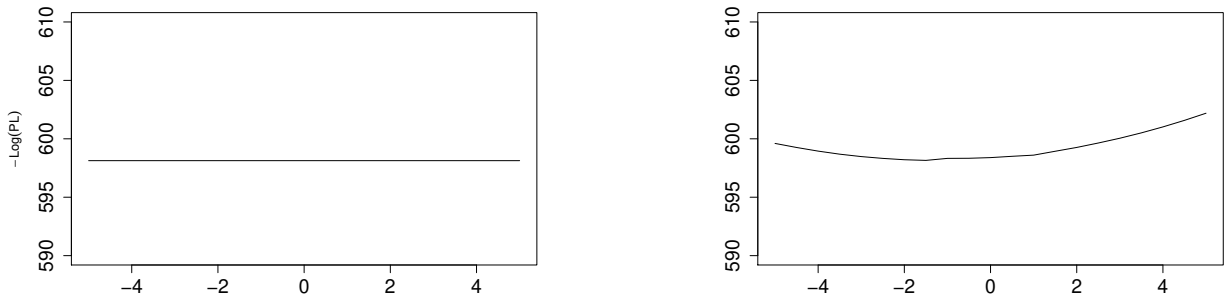
In our model, however, intercept term β_0, is not our object of interest. Thus, this issue would not impact the application of our model.

5.5 Simulation

We conduct simulations to validate if the suggested method could estimate the true covariate effects on the marginal mean. Here we consider two types of data generating mechanisms: (1) Data are generated from the marginal mean hazard rate model; and (2) Data are generated from the standard conditional cure rate model.

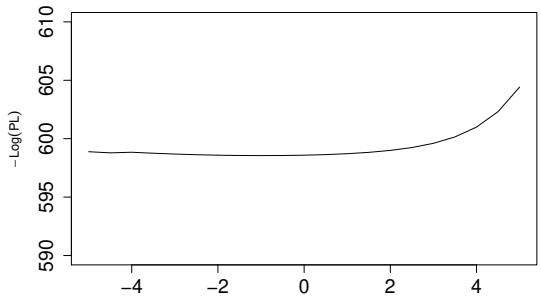
5.5.1 Data generated from the marginal model

In the first simulation, the survival time $t_i (i = 1, 2, ..., n)$ is generated from the marginal cure rate model with the true $E[h_M(t_i)] = \exp\{-1.5 - 1.25X_{i1} - 0.75X_{i2}\}$, where $X_{i1} \sim N(0, 1)$

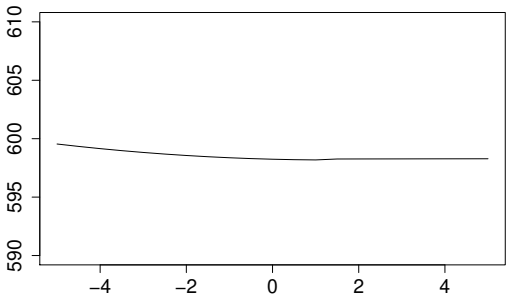


(a) β_0

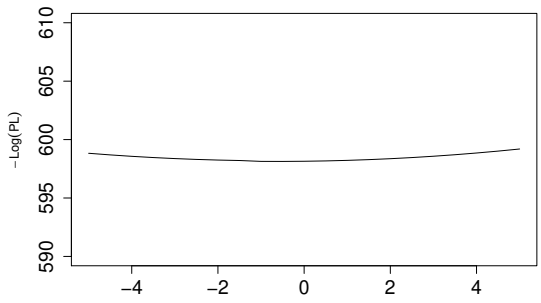
(b) β_1



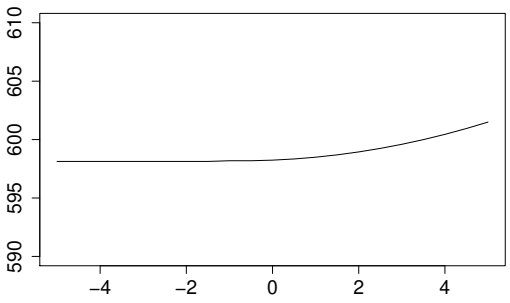
(c) β_2



(d) γ_0



(e) γ_1



(f) γ_2

Figure 5.1: Likelihood profiles for the marginal mean and uncured fraction parameters in the estimation, where $n = 400$ and $\lambda_c = 0.001$

and $X_{i2} \sim U(0,1)$. To generate survival time for the uncured group, we need to derive the hazard function as well as density of it by taking use of marginal mean hazard $E[h_M(t_i)]$. We assume a true Weibull baseline hazard function for the uncured group while using Bernstein polynomials to fit this baseline hazard. Then the conditional hazard function or uncured hazard function for data generation is

$$\begin{aligned} h_u(t_i) &= \alpha t_i^{\alpha-1}\Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\pi_i\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha} \\ &= 0.8 \times t_i^{-0.2}\Big[\frac{\exp\{-1.5-1.25X_{1i}-0.75X_{2i}\}}{0.8\times\pi_i\times\Gamma(3/4)}\Big]^{0.8}, \end{aligned}$$

where $\text{logit}(\pi_i) = 1.5-0.5X_{1i}-0.75X_{2i}$. Based on the above setting, the data generating procedure is the same as the one we described in Section 4.3.1. The censoring time are generated from an exponential distribution with rate λ_c taking values on 0.1, 0.01, 0.001, representing mild, intermediate and heavy censoring. According to Appendix A.3, the true censoring rate for the uncured population is 29.80%, 7.62% and 4.01% respectively. Simulations are repeated 1000 times for sample sizes from 200 to 800.

The estimation result of mild censoring using both true model and Bernstein polynomial methods are presented in Table 5.1. We observe that in the setting of mild censoring, Bernstein methods are as good as the true Weibull assumption when sample size is more than 200. We also see that the standard error and bias of the marginal parameters β_1 and β_2 decreases as the sample size increases. This empirical result indicates our semi-parametric should have asymptotic consistency property and is robust with respect to the function of uncured baseline hazard.

The results of intermediate censoring are presented in Table 5.2. We observe that in the setting of intermediate censoring, estimation results using Bernstein methods are improved and as good as true Weibull assumption method when sample size reaches 600. We also see that the standard error and bias of the marginal parameters β_1 and β_2 decreases as the sample size increases.

However, the results of heavy censoring, presented in Table 5.3, indicates that the esti-

mation of marginal parameters deviates from the true values when censoring rate is high, especially when the sample size is 200 or 400. The estimation is improved when sample size reached 800. Thus, reseacher should demand a large sample size in real practice when heavy censoring is present in the data.

Table 5.1: Comparison of Bernstein baseline hazard and Weibull baseline hazard for marginalized mean hazard rate model with 1000 simulations and mild censoring rate ($\lambda_c = 0.001$)

		Weibull baseline hazard			Bernstein baseline hazard		
	True	Esti.	SE	CP	Esti.	SE	CP
				$n = 200$			
β_0^*	-1.5	-1.491	0.276	0.951	10.277	0.636	0.000
β_1^*	-1.25	-1.253	0.151	0.937	-1.252	0.082	0.970
β_2^*	-0.75	-0.735	0.466	0.944	-0.735	0.288	0.965
ξ_0^*	1.5	1.572	0.627	0.951	1.515	0.255	0.960
ξ_1^*	-0.5	-0.515	0.309	0.950	-0.514	0.133	0.935
ξ_2^*	-0.75	-0.821	0.933	0.952	-0.777	0.399	0.945
				$n = 400$			
β_0^*	-1.5	-1.490	0.194	0.960	10.262	0.592	0.000
β_1^*	-1.25	-1.251	0.105	0.946	-1.247	0.082	0.949
β_2^*	-0.75	-0.749	0.326	0.948	-0.742	0.274	0.954
ξ_0^*	1.5	1.532	0.421	0.962	1.517	0.247	0.962
ξ_1^*	-0.5	-0.509	0.210	0.952	-0.518	0.119	0.941
ξ_2^*	-0.75	-0.777	0.627	0.966	-0.742	0.387	0.956
				$n = 600$			
β_0^*	-1.5	-1.492	0.157	0.947	9.912	0.363	0.000
β_1^*	-1.25	-1.249	0.085	0.941	-1.246	0.069	0.950
β_2^*	-0.75	-0.747	0.264	0.962	-0.761	0.229	0.950
ξ_0^*	1.5	1.547	0.341	0.962	1.515	0.180	0.940
ξ_1^*	-0.5	-0.508	0.170	0.934	-0.508	0.103	0.960
ξ_2^*	-0.75	-0.795	0.508	0.958	-0.753	0.299	0.955
				$n = 800$			
β_0^*	-1.5	-1.493	0.112	0.953	9.792	0.157	0.000
β_1^*	-1.25	-1.253	0.058	0.945	-1.246	0.060	0.946
β_2^*	-0.75	-0.761	0.194	0.951	-0.754	0.210	0.952
ξ_0^*	1.5	1.497	0.177	0.942	1.513	0.142	0.942
ξ_1^*	-0.5	-0.503	0.088	0.934	-0.507	0.090	0.944
ξ_2^*	-0.75	-0.741	0.292	0.942	-0.748	0.212	0.948

CP: Coverage Probability
SE: Standard Error

Table 5.2: Comparison of Bernstein baseline hazard and Weibull baseline hazard for marginalized mean hazard rate model with 1000 simulations and intermediate censoring rate ($\lambda_c = 0.01$)

		Weibull baseline hazard			Bernstein baseline hazard		
Para.	True	Esti.	SE	CP	Esti.	SE	CP
				$n = 200$			
β_0^*	-1.5	-1.473	0.234	0.944	9.854	0.282	0.000
β_1^*	-1.25	-1.252	0.125	0.943	-1.233	0.126	0.955
β_2^*	-0.75	-0.776	0.409	0.943	-0.764	0.429	0.950
ξ_0^*	1.5	1.521	0.393	0.962	1.515	0.255	0.960
ξ_1^*	-0.5	-0.509	0.211	0.957	-0.585	0.939	0.995
ξ_2^*	-0.75	-0.741	0.658	0.957	-0.685	0.419	0.940
				$n = 400$			
β_0^*	-1.5	-1.492	0.164	0.947	10.177	0.614	0.000
β_1^*	-1.25	-1.250	0.088	0.951	-1.231	0.080	0.955
β_2^*	-0.75	-0.754	0.287	0.932	-0.730	0.281	0.945
ξ_0^*	1.5	1.509	0.275	0.946	1.460	0.259	0.950
ξ_1^*	-0.5	-0.506	0.147	0.954	-0.522	0.163	0.935
ξ_2^*	-0.75	-0.739	0.459	0.948	-0.738	0.421	0.945
				$n = 600$			
β_0^*	-1.5	-1.491	0.134	0.952	10.013	0.181	0.000
β_1^*	-1.25	-1.249	0.072	0.943	-1.235	0.068	0.935
β_2^*	-0.75	-0.764	0.264	0.949	-0.758	0.224	0.94
ξ_0^*	1.5	1.511	0.223	0.964	1.509	0.233	0.945
ξ_1^*	-0.5	-0.511	0.119	0.955	-0.518	0.123	0.960
ξ_2^*	-0.75	-0.770	0.371	0.955	-0.772	0.399	0.950
				$n = 800$			
β_0^*	-1.5	-1.504	0.116	0.961	9.870	0.125	0.000
β_1^*	-1.25	-1.247	0.062	0.942	-1.239	0.062	0.945
β_2^*	-0.75	-0.738	0.202	0.949	-0.744	0.215	0.955
ξ_0^*	1.5	1.503	0.193	0.941	1.492	0.198	0.950
ξ_1^*	-0.5	-0.502	0.103	0.951	-0.526	0.103	0.935
ξ_2^*	-0.75	-0.727	0.322	0.954	-0.745	0.338	0.945

CP: Coverage Probability
SE: Standard Error

Table 5.3: Comparison of Bernstein baseline hazard and Weibull baseline hazard for marginalized mean hazard rate model with 1000 simulations and heavy censoring rate ($\lambda_c = 0.1$)

		Weibull baseline hazard			Bernstein baseline hazard		
	True	Esti.	SE	CP	Esti.	SE	CP
				$n = 200$			
β_0^*	-1.5	-1.463	0.293	0.938	6.166	0.603	0.000
β_1^*	-1.25	-1.241	0.169	0.933	-1.145	0.161	0.895
β_2^*	-0.75	-0.752	0.516	0.941	-0.670	0.510	0.945
ξ_0^*	1.5	1.632	0.804	0.944	3.865	22.255	0.990
ξ_1^*	-0.5	-0.489	0.443	0.930	-2.157	13.230	0.985
ξ_2^*	-0.75	-0.695	1.317	0.965	2.109	42.568	0.990
				$n = 400$			
β_0^*	-1.5	-1.480	0.207	0.935	7.192	1.315	0.000
β_1^*	-1.25	-1.250	0.121	0.942	-1.157	0.100	0.840
β_2^*	-0.75	-0.760	0.364	0.946	-0.683	0.335	0.950
ξ_0^*	1.5	1.591	0.538	0.946	1.357	0.424	0.945
ξ_1^*	-0.5	-0.483	0.308	0.952	-0.645	0.284	0.915
ξ_2^*	-0.75	-0.753	0.850	0.954	-0.769	0.759	0.965
				$n = 600$			
β_0^*	-1.5	-1.489	0.169	0.950	7.157	0.868	0.000
β_1^*	-1.25	-1.246	0.098	0.948	-1.185	0.092	0.880
β_2^*	-0.75	-0.749	0.297	0.952	-0.696	0.323	0.950
ξ_0^*	1.5	1.567	0.417	0.955	1.334	0.372	0.945
ξ_1^*	-0.5	-0.492	0.245	0.939	-0.640	0.221	0.895
ξ_2^*	-0.75	-0.777	0.665	0.956	-0.797	0.593	0.960
				$n = 800$			
β_0^*	-1.5	-1.481	0.145	0.938	6.990	0.735	0.000
β_1^*	-1.25	-1.247	0.085	0.948	-1.212	0.070	0.915
β_2^*	-0.75	-0.767	0.256	0.948	-0.698	0.248	0.940
ξ_0^*	1.5	1.543	0.349	0.963	1.297	0.271	0.880
ξ_1^*	-0.5	-0.502	0.207	0.942	-0.601	0.182	0.925
ξ_2^*	-0.75	-0.790	0.558	0.960	-0.681	0.472	0.940

CP: Coverage Probability
SE: Standard Error

5.5.2 Data generated from the conditional model

We conducted simulations to evaluate the performance of nonparametric marginal mean hazard rate method with the Average Predicted Value (APV) approach proposed by Albert et al. (2014). The survival time t_i is generated from the conditional cure rate model with

$$\begin{aligned} h_u(t_i) &= h_{u0}(t_i) \exp\left\{\boldsymbol{\eta}' \boldsymbol{x_i}\right\} \\ &= 0.75 \times 0.1 \times t_i^{-0.25} \times \exp\{-4 + X_{i1} - 2X_{i2}\} \end{aligned}$$

where X_1 is a binary exposure covariate taking the value of 1 if subject is exposed, and 0 otherwise. We assume X_{i1} to be a balanced categorical variable with $P(X_{i1} = 1) = 0.5$. For X_{i2}, we consider using a standard normal distribution or a uniform distribution. The uncure fraction is assumed to be $\text{logit}(\pi_i) = 1 + 2X_{i1} + 1.5X_{i2}$. Censoring time is generated from an exponential distribution with $\lambda_c = 0.0002$.

The simulations are replicated 1000 times for sample sizes from 200 to 800. We implement the simulations by using the Quasi-Newton nonlinear optimization embedded in SAS 9.4 NLMIXED procedure (SAS Institue, Cary, NC, USA).

The simulation results presented in Table 5.4 indicate that both nonparametric Bernstein method and parametric Weibull method perform well in estimating the overall exposure effect on marginal mean hazard. As sample sizes increases, the estimated mean ratio is closer to the true mean ratio and the corresponding standard errors decreases.

5.6 Application

In this section, we apply the proposed Bernstein marginal mean hazard rate model suggested above to the liver cancer data discussed in Section 4.4. The liver cancer data are collected from the Surveillance, Epidemiology, and End Results (SEER) Program in Connecticut. Our dataset contains a total of 2362 patients aged from 10 to 96 who were diagnosed with liver cancer between 1975 and 2016. The event in this study is the death of liver cancer. Our primary objective in this study is to evaluate the covariate effects such as age or surgery

Table 5.4: Simulation results for mean ratio (exposed vs nonexposed) when data generated from the conditional model with two covariates and $\lambda_c = 0.0002$

	$X_2 \sim U(0,1)$				$X_2 \sim N(0,1)$			
	Weibull		Bernstein		Weibull		Bernstein	
n	Esti.	SE	Esti.	SE	Esti.	SE	Esti.	SE
200	4.884	0.635	4.898	0.521	4.695	0.473	4.678	0.512
400	4.730	0.590	4.890	0.482	4.740	0.460	4.671	0.487
600	4.778	0.524	4.873	0.465	4.741	0.449	4.762	0.463
800	4.850	0.442	4.856	0.421	4.772	0.421	4.802	0.438
True θ_R	4.845				4.781			

treatment on the overall mean hazard of liver cancer patients, that is, the overall risk of death. By comparing the parametric and nonparametric methods, we should expect similar parameter estimations.

As shown in Section 4.4, the follow-up time for the liver cancer study is long enough to observe the cured patients. As long-term survivors are justified, we apply the proposed methods to the data with covariates of patient's age at diagnosis (Age), surgery or not (SUR) and number of lymph nodes (NLN). The marginal mean hazard rate and uncured fraction are related to covariates as follows,

$$E[h_M(t_i)] = \exp\left\{\beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{SUR}_i + \beta_3 \text{NLN}_i\right\} \quad \text{and}$$
$$\text{logit}(\pi_i) = \xi_0 + \xi_1 \text{AGE}_i + \xi_2 \text{SUR}_i + \xi_3 \text{NLN}_i$$

The estimates and the standard errors of parameters of interest are presented in Table 5.5. As seen, covariates Age and Number of lymph nodes have positive effects on the overall mean hazard rate ($\hat{\beta}_1 = 0.108$, p-value 0.002; $\hat{\beta}_3 = 0.247$, p-value < 0.001), which indicates that the patients diagnosed at older age with higher number of lymph nodes might expose to a higher risk of death. Surgery, however, is negatively associated with the marginal mean hazard rate ($\hat{\beta}_2 = -0.937$, p-value < 0.001), meaning that the liver cancer patients who received the surgery would expect to have a lower risk of death. The comparison results that

Table 5.5: The parameter estimations for the liver cancer patients using the marginalized mean hazard rate model with Weibull baseline hazard and Bernstein baseline assumption

			Weibull baseline hazard			Bernstein baseline hazard		
	Covariate	Para.	Esti.	SE	p-value	Esti.	SE	p-value
Mean hazard	Age	β_1	0.117	0.035	< 0.001	0.108	0.036	0.002
	Surgery	β_2	-0.948	0.135	< 0.001	-0.937	0.208	< 0.001
	Lymph Nodes	β_3	0.249	0.029	< 0.001	0.247	0.032	< 0.001
Uncure fraction	Age	ξ_1	0.201	0.052	< 0.001	0.153	0.048	0.002
	Surgery	ξ_2	-1.306	0.373	< 0.001	-1.303	0.324	< 0.001
	Lymph Nodes	ξ_3	0.395	0.058	< 0.001	0.390	0.079	< 0.001

two methods give similar estimations of marginal parameters are as we expected, indicating the above interpretation of covariates effects on the overall population is reliable.

5.7 Discussion

In the above sections, we employ the Bernstein polynomials to relax the parametric baseline hazard assumption for the marginal mean hazard rate model. Compared with other nonparametric techniques such as M or I splines, Bernstein polynomials enjoy the advantage of preserving shape and supporting fast implementation. As the equation of Bernstein polynomials is similar to the kernel of Beta distribution, we could conveniently calculate the $E[h_{u0}(x)]$, existing in the marginal likelihood function, by Beta approximation. However, using splines such as M and I splines (Ramsay et al., 1988) will make the computation of $E[h_{u0}(t)]$ challenging. Thus it is hard to estimate the likelihood function of the marginal mean hazard rate model, and those splines are not used.

In addition, there is no need to decide the spline knots for Bernstein polynomials. The simulation results indicate that estimated parameters are consistent. Furthermore, the Bernstein estimation is as efficient as the Weibull estimation even though the true data are generated from the Weibull baseline hazard function. Compared with the previous estimates, the

application results indicate that the proposed semi-parametric marginal model works well in the real practice.

Chapter 6

Marginal mean hazard rate model with high-dimensional data

6.1 Motivation

This extension of our marginal mean hazard rate model to high-dimensional data is motivated by TCGA (The Cancer Genome Atlas) breast cancer data. TCGA provides high-quality genomic information which can be used to identify the abnormalities in DNA that are associated with the hazardous cells. In our study, we are interested in identifying the important microarrays in the TCGA microarray data that are significantly related to survival outcome of the breast cancer patients. However, in this survival dateset we observe a certain proportion of censoring after a long period of follow-up, indicating the existence of long-term survivors. Therefore, it is appropriate to use cure rate models to anlyze such data.

Few studies have addressed variable selection associated with the censored surviving data, especially in the high-dimensional covariates with small sample size setting. Tibshirani (1997) firstly imposed a LASSO penalty on the partial likelihood function for Cox proportional hazard rate model in the low-dimensional setting. Later he developed a quadratic programming technique that could overcome the diffculty of computation for optimizing the

above penalized likelihood function in high-dimensional setting (Tibshirani, 2009). Similar studies include Gui and Li (2005), using both quadratic programming and LARS algorithm to do estimation for LASSO-type penalized Cox model under the high-dimensional setting. However, these studies are limited to the homogeneous survival data with only uncured population. Some other methods involving variable selection for cure rate model are only investigated for the low-dimensional data (Liu et al., 2012; Masud et al., 2018). An exception is Fan et al. (2017), proposing a penalization method for variable selection under the standard cure rate model, where the same covariate structure and coefficients on both survival rate and cure rate are assumed. While this could promote the structure effects, it is not an approriate model if covariates have different bilogocial process on affecting the cure rate and survival rate for the uncured subpopulation. Furthermore, there is no test procedure available for examining the similarity in the covariate structure.

The goals of our study inlcude: (1) identify the key microarrays that are associated with overall hazard rate of breast cancer patients in the high-dimensional setting, and (2) propose a rubost estimation method without the need of correctly specifiying the unknown relation between cure rate and covariates. To achieve these goals, we extend the proposed marginal cure rate model to the high-dimensional setting. In this high-dimensional setting, we further assume a working random distribution for cure fraction. This randomization technique could help reduce the number of parameters as well as avoid the issue of misspecification on cure fraction. We propose a penalization method for conducting variable selection for the marginal cure rate model in the high-dimensional data setting. The penalized marginal likelihood could be estimated by linear quadrature-cordinate decent method. To the best of our knowledge, this is the first paper that studying the variable selection for marginalized model in the high-dimensional setting.

6.2 Marginal mean hazard rate model with random cure fraction

As discussed in Chapter 4, we relate marginal mean hazard rate $E[h_M(t_i)]$ to covariates directly by a log link function, and assume a Weibull proportional hazard for the uncured subpopulation. Specifically,

$$\begin{aligned} E[h_M(t_i)] &= e^{\boldsymbol{\beta' x_i}} \\ h_u(t_i) &= h_{u0}(t)e^{\mu_i} \\ h_{u0}(t_i) &= \alpha\lambda t_i^{\alpha-1}, \end{aligned} \tag{6.1}$$

where μ_i is not necessarily a linear function of covariates, α and λ are scale and shape parameters of Weibull distribution, respectively.

It is common to see in the literature that covariates are related to uncure fraction π through a logit link function. However, it is very difficult to validate this specification in practice, and the high-dimensional covariates may cause an issue of model fitting due to the curse of dimensionality. In this study, covariate effects on the cure fraction are not the primary of interest. We could assume a random distribution rather than using a logit link function for the cure fraction. This assumption can avoid the misspecification if the true link function is unknown for us.

Based on Equation 4.3 and the above assumption, we can further have the conditional hazard rate as

$$\begin{aligned} h_u(t_i) &= \alpha t_i^{\alpha-1}\Big[\frac{E[h_M(t_i)]}{\alpha\pi_i\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha} \\ &= \alpha t_i^{\alpha-1}\Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\pi\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha}, \quad \text{where } \pi \text{ is a random variable,} \end{aligned} \tag{6.2}$$

which can be embedded in the likelihood function of a standard cure rate model. The associated marginal likelihood is then obtained by intergrating out π from this likelihood function.

With the random assumption of π, the covariates are no longer related to the uncured fraction. However, the marginal mean hazard rate model is still needed when the research goal is interpreting the covariate effect on the marginal response. While the standard cure rate model with the random π is good to interpret the covariate effects on $h_u(t_i)$, it is hard to obtain the interpretation of covariate effects on the marginal outcome. For example,

$$\begin{aligned} h_M(t_i|\boldsymbol{w_i}) &= E_\pi\left[\frac{\pi f_u(t_i)}{1-\pi+\pi S_u(t_i)}\right] \\ &= \int_0^1 \frac{\pi h_u(t_i) S_u(t_i)}{1-\pi+\pi S_u(t_i)} f(\pi) d\pi \\ &= \int_0^1 \frac{\pi h_{u0}(t_i) \exp\{\boldsymbol{\eta' w_i}\}\left[S_{u_0}(t_i)\right]^{-\exp\{\boldsymbol{\eta' w_i}\}}}{1-\pi+\pi\left[S_{u_0}(t_i)\right]^{-\exp\{\boldsymbol{\eta' w_i}\}}} f(\pi) d\pi \end{aligned}$$

Clearly, the effect of covariates $\boldsymbol{w}_i$ on $h_M(t_i)$ is changing over different values of $\boldsymbol{w}_i$ and t_i, which makes the interpretation difficult from the marginal perspective. Thus, the marginal mean hazard rate model, relating the marginal response directly to the covariates, is still needed.

6.3 Penalized likelihood function

6.3.1 Marginal likelihood function

Suppose we have independently observed data $(t_i, \delta_i, \boldsymbol{x_i})$, where t_i is the observed event time, δ_i is the censoring indicator with a value 1 if t_i is not censored and 0 otherwise. $\boldsymbol{x_i}$ is the covariate vector related to the marginal hazard rate. Then, marginal likelihood $\mathcal{L}(\alpha, \theta, \boldsymbol{\beta})$ by intergrating out π when $\delta_i = 1$ is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \alpha, \theta|t_i, \delta_i = 1) &= \prod_{\delta_i=1} E_\pi[\mathcal{L}(\boldsymbol{\beta}, \alpha, \theta|t_i, \pi)] \\ &= \prod_{\delta_i=1} \alpha\Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\Gamma(2-\frac{1}{\alpha})}\Big]^\alpha t_i^{\alpha-1} E_\pi\left[\pi^{-\alpha+1}\exp\Big\{-t_i^\alpha\Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha\Gamma(2-\frac{1}{\alpha})}\Big]^\alpha \pi^{-\alpha}\Big\}\right] \end{aligned}$$

Let $b_i = t_i^{\alpha}\left[\frac{e^{\boldsymbol{\beta}'\boldsymbol{x_i}}}{\alpha\Gamma(2-\frac{1}{\alpha})}\right]^{\alpha}$, and $u = \pi^{-\alpha}$ with range $[1, +\infty)$ as $\pi \in [0, 1]$. then

$$\mathcal{L}(\boldsymbol{\beta}, \alpha, \theta | t_i, \delta_i = 1) = \prod_{\delta_i=1} \frac{\alpha b_i}{t_i} E_u u^{1-\frac{1}{\alpha}} \exp\{-b_i u\} \tag{6.3}$$

We assume u or $\pi^{-\alpha}$ follows a two-parameter exponential distribution with the scale parameter θ and the location parameter of 1, i.e, this pdf is $f(u) = \theta \exp\{-\theta(u-1)\}$. By assuming a random π, we could reduce about half the amount of parameters in the high-dimensional setting and avoid the misspecification if the true link function is unknown. We should choose a random distribution that could represent the uncertainty of uncure fraction. There are some advantages by using the two-parameter exponential distribution. First, the support is matched with $\pi^{-\alpha}$. Second, the exponential form of this distribution will be beneficial to calculate the likelihood function, which could be seen in the Equations 6.3 and 6.4. Finally, this random assumption could account for the uncertainty of cure fraction over different individuals. Then the integration in the above marginal likelihood function can be done as

$$\begin{aligned} E_u u^{1-\frac{1}{\alpha}} \exp\{-b_i u\} &= \int_1^{\infty} u^{1-\frac{1}{\alpha}} \exp\{-b_i u\} \theta \exp\{-\theta(u-1)\} du \\ &= \frac{\theta e^{-b_i}}{\theta + b_i} \int_1^{\infty} u^{1-\frac{1}{\alpha}} (\theta + b_i) \exp\{-(\theta + b_i)(u-1)\} du \\ &= \frac{\theta e^{-b_i}}{\theta + b_i} M(1 - \frac{1}{\alpha}), \end{aligned} \tag{6.4}$$

where $M(1-\frac{1}{\alpha})$ is the $(1-\frac{1}{\alpha})^{\text{th}}$ moment of exponential distribution with the scale parameter

$\theta + b_i$ and the location parameter of 1. The marginal log-likelihood function for $\delta_i = 1$is then

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \alpha, \theta | t, \delta_i = 1) &= \sum_{\delta_i=1} \frac{\alpha\theta b_i e^{-b_i}}{(\theta + b_i)t_i} M(1 - \frac{1}{\alpha}) \\
&= \sum_{\delta_i=1} \log(\alpha) + \log(\theta) + \alpha\Big[\boldsymbol{\beta'x_i} - \log(\alpha) - \log\Gamma\Big(2 - \frac{1}{\alpha}\Big)\Big] - t_i^{\alpha}\Big[\frac{e^{\boldsymbol{\beta'x_i}}}{\alpha\Gamma(2 - \frac{1}{\alpha})}\Big]^{\alpha} \\
&- \log\Big\{\theta + t_i^{\alpha}\Big[\frac{e^{\boldsymbol{\beta'x_i}}}{\alpha\Gamma(2 - \frac{1}{\alpha})}\Big]^{\alpha}\Big\} - \log(t_i) + \log\Big\{M\Big(1 - \frac{1}{\alpha}\Big)\Big\}
\end{aligned} \tag{6.5}$$

Likewise, for $\delta_i = 0$, the marginal likelihood function is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \alpha, \theta | t_i, \delta_i = 0) &= \prod_{\delta_i=0} E_{\pi}[\mathcal{L}(\boldsymbol{\beta}, \alpha, \theta | t_i, \pi)] \\
&= \prod_{\delta_i=0} \Bigg\{ E[1 - \pi] + E\Bigg[\pi \exp\Big\{ - t^{\alpha}\Big[\frac{e^{\boldsymbol{\beta'x_i}}}{\alpha\Gamma(2 - \frac{1}{\alpha})}\Big]^{\alpha}\pi^{-\alpha}\Big\}\Bigg]\Bigg\}
\end{aligned}$$

Assuming $b_i = t_i^{\alpha}[\frac{e^{\boldsymbol{\beta'x_i}}}{\alpha\Gamma(2-\frac{1}{\alpha})}]^{\alpha}$, and $u = \pi^{-\alpha}$ with range $[1, +\infty)$ as $\pi \in [0, 1]$, then

$$\mathcal{L}(\beta, \alpha, \theta | t_i, \delta_i = 0) = \prod_{\delta_i=0} \Big\{E[1 - \pi] + \frac{\theta e^{-b_i}}{\theta + b_i} M(\alpha)\Big\},$$

where $M(\alpha)$ is the α^{th} moment of exponential distribution with the scale parameter $(\theta + b_i)$ and the location parameter of 1. And

$$\begin{aligned}
E[\pi] &= E[u^{-\frac{1}{\alpha}}] \\
&= \int_1^{\infty} u^{-\frac{1}{\alpha}} \theta e^{-\theta(u-1)} du \\
&= \theta e^{\theta} \int_1^{\infty} u^{-\frac{1}{\alpha}} e^{-\theta u} du \quad \text{let } \nu = \theta u, \\
&= \theta^{\frac{1}{\alpha}} e^{\theta} \int_{\theta}^{\infty} \nu^{-\frac{1}{\alpha}} e^{-\nu} d\nu \\
&= \theta^{\frac{1}{\alpha}} e^{\theta} \Gamma(1 - \frac{1}{\alpha}, \theta),
\end{aligned}$$

where $\Gamma(1 - \frac{1}{\alpha}, \theta)$ is the incomplete upper Gamma fuction. Then marginal log-likelihood for

censored individuals is

$$\begin{aligned}\ell(\beta,\alpha,\theta|t_i,\delta_i=0) &= \sum_{\delta_i=0} E_\pi\big[\ell(\beta,\alpha,\theta|t_i,\pi)\big] \\ &= \sum_{\delta_i=0}\Big\{\log\Big(1-\theta^{\frac{1}{\alpha}}e^{\theta}\Gamma(1-\frac{1}{\alpha},\theta)+\frac{\theta e^{-b_i}}{\theta+b_i}M(\alpha)\Big)\Big\}\end{aligned} \tag{6.6}$$

Finally, the marginal log-likehood function for all subjects is then given as (i.e. Equation 6.5 and 6.6),

$$\begin{aligned}\ell(\boldsymbol{\beta},\alpha,\theta|t_i) = \sum_{\delta_i=1}\Bigg\{ &\log(\alpha)+\log(\theta)+\alpha\Big[\boldsymbol{\beta'x_i}-\log(\alpha)-\log\Gamma\Big(2-\frac{1}{\alpha}\Big)\Big]-t_i^{\alpha}\Big[\frac{e^{\boldsymbol{\beta'x_i}}}{\alpha\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha} \\ &-\log\Big\{\theta+t_i^{\alpha}\Big[\frac{e^{\boldsymbol{\beta'x_i}}}{\alpha\Gamma(2-\frac{1}{\alpha})}\Big]^{\alpha}\Big\}-\log(t_i)+\log\Big\{M\Big(1-\frac{1}{\alpha}\Big)\Big\}\Bigg\} \\ &+\sum_{\delta_i=0}\Bigg\{\log\Big(1-\theta^{\frac{1}{\alpha}}e^{\theta}\Gamma(1-\frac{1}{\alpha},\theta)+\frac{\theta e^{-b_i}}{\theta+b_i}M(\alpha)\Big)\Bigg\}\end{aligned}$$

6.3.2 Penalization and estimating algorithm

By adopting the LASSO approach, we impose a ℓ_1 penalty on the marginal log-likelihood function given α and θ in order to induce sparsity to high-dimensional covariates for variable selection (Tibshirani, 1997), that is,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{-\ell(\boldsymbol{\beta},\alpha,\theta)\Big\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s$$

where p is the number of covariates and α is the shape parameter of Weibull baseline, and θ is the parameter characterizing the random variable π. User-specified parameter s can be determined by cross-validation method.

We employ the techniques of quadratic approximation (Tibshirani, 1997) and cyclical coordinate descent algorithm (Friedman et al., 2010) to estimate $\boldsymbol{\beta}$. Quadratic techinique can approximate the complicated marginal log-likelihood function by using a quadratic form derived from the second-order Taylor expansion of that function. Since this quadrature

approximation has the same form as the squared error of general linear model, we could easily obtain the optimizer by employing existing R package such as *lars* or *glmnet*. The later package uses the cyclical coordinate descent algorithm which is highly efficient to compute the solution path for the LASSO or ridge regression. In this study, we use the *glmnet* package to do the parameter estimation after giving the form of the quadratic approximation under our framework.

To give specific details on this estimation, we briefly describe the procedure below. The quadratic method to approximate the likelihood function is formulated as follows: Suppose X is the design matrix of covariates and $\boldsymbol{\zeta} = X\boldsymbol{\beta}$, define $\boldsymbol{\mu} = \partial\ell/\partial\boldsymbol{\zeta}$, $A = -\partial^2\ell/\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T$ and $\boldsymbol{z} = \boldsymbol{\zeta} + A^-\boldsymbol{\mu}$, then the quadrature approximation for the above $\ell(\boldsymbol{\beta})$ after updating α and θ by Taylor expansion has the form

$$\ell(\boldsymbol{\beta}|\alpha,\theta) \approx (\boldsymbol{z} - X\boldsymbol{\beta})^T A(\boldsymbol{z} - X\boldsymbol{\beta}) \tag{6.7}$$

Even though A^- is not unique, Gui and Li (2005) pointed out that Equation 6.7 is invariant to the choice of A^- that satisfies $AA^-A = A$ as long as $rank(A) = n - 1$. Set $Q = A^{1/2}$, $\tilde{\boldsymbol{z}} = Q\boldsymbol{z}$, $\tilde{X} = QX$, then above equation could be denoted as

$$\ell(\boldsymbol{\beta}|\alpha,\theta) \approx (\tilde{\boldsymbol{z}} - \tilde{X}\boldsymbol{\beta})^T(\tilde{\boldsymbol{z}} - \tilde{X}\boldsymbol{\beta})$$

Then the modified iterative procedure of Tibshirani (1997) given below are used to obtain $\hat{\boldsymbol{\beta}}$:

(1) Initialize $\hat{\boldsymbol{\beta}} = \beta^{(0)}$, such as $\beta^{(0)} = 0$.

(2) Estimate the nuissance parameter θ and α.

(3) Compute $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, A, $\boldsymbol{z}$, Q, $\tilde{\boldsymbol{z}}$, $\tilde{X}$ based on current value of θ, α and $\boldsymbol{\beta}$.

(4) Update $\hat{\boldsymbol{\beta}}$ by minimizing $(\tilde{\boldsymbol{z}} - \tilde{X}\boldsymbol{\beta})^T(\tilde{\boldsymbol{z}} - \tilde{X}\boldsymbol{\beta})$ subject to $\sum_{j=1}^{p}|\beta_j| \leq s$ based on the cyclical coordinate descent algorithm.

(5) Repeat (2),(3),(4) until convergency of $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$.

The above estimation algorithm for marginal mean hazard rate lasso model could also

be extended to the elastic net or ridge by modifying the step (4). We could update $\hat{\boldsymbol{\beta}}$ by minimizing $(\tilde{\boldsymbol{z}}-\tilde{X}\boldsymbol{\beta})^T(\tilde{\boldsymbol{z}}-\tilde{X}\boldsymbol{\beta})$ subject to $\sum_{j=1}^{p}\beta_j^2 \leq s$ for ridge and subject to $\sum_{j=1}^{p}\lambda|\beta_j|+\sum_{j=1}^{p}(1-\lambda)\beta_j^2 \leq s$ for elastic net. R package *glmnet* could be used directly for the update.

6.4 Numerical study

6.4.1 Simulation

In this simulation, the survival time $t_i (i = 1, 2, ..., n)$ is generated from the marginal cure rate model with the true $E[h_M(t)] = \exp\{\boldsymbol{\beta'x_i}\}$, where total number of covariates in $\boldsymbol{x}_i$ is p set as 100 or 300, and these covariates are generated from a multivariate normal distribution with mean zero and correlation ρ and variance of 1. We do consider some correlation between covariates, $\boldsymbol{X}$ is multivariate normal with $\text{Cov}(X_j, X_k) = \rho^{|j-k|}$, where $\rho = 0, 0.2, 0.5, 0.7$. We are assuming 6 covariates are related to the outcome (i.e., positive) and the rest are unrelated (i.e., negative). Without loss of generality, the first 5 covariates are true positives and their corresponding parameters are $(1, -1.5, 2, 1.75, -1.25)$, and the intercept parameter is -1. For the rest of covariates, the parameters are set to zero. We consider parameter $\alpha^* = 0.9$ or $\alpha^* = 1.1$ for the Weibull baseline hazard. We also assume $\pi^{-\alpha^*} \sim \text{Exp}(\theta^*, 1)$, where $\theta^* = 6$ in order to achieve a non-negligible cure fraction in the data. The censoring time for each individual is generated from the exponential distribution with rate λ_c: (1) Heavy censoring with $\lambda_c = 0.02$; (2) Intermediate censoring with $\lambda_c = 0.002$; (3) Mild censoring with $\lambda_c = 0.0002$.

Simulations are replicated 500 times for sample size $n = 200$. Simulation results from Table 6.1 and 6.2 show increasing biases as the correlation between covairates are greater. But the false positive rate and false negative rate are relatively low.

Figure 6.1 and 6.2 illustrate the selection rate of each covariate in the simulated data under the low censoring and the high censoring setting, respectively. As shown, the rates of selecting correct covariate are roughly between 0.9 and 1 when the correlation is small to moderate (i.e., $\rho = 0$, $\rho = 0.2$, $\rho = 0.5$, see Figure 6.1(a),(b),(c), for example). But when

Table 6.1: Simulation results when $\alpha^* = 1.1$ and $\theta^* = 6$

	$n = 200, p = 100$			$n = 200, p = 300$		
Corr. ρ	Model Error	FPR	FNR	Model Error	FPR	FNR
Mild Censoring($\lambda_c = 0.0002$)						
0	0.541	0.037	0.000	0.667	0.012	0.000
0.2	0.569	0.039	0.000	0.735	0.015	0.000
0.5	0.713	0.056	0.000	0.981	0.022	0.000
0.7	0.885	0.080	0.000	1.446	0.034	0.025
Moderate Censoring($\lambda_c = 0.002$)						
0	0.608	0.041	0.000	0.779	0.015	0.000
0.2	0.737	0.061	0.000	1.064	0.024	0.003
0.5	0.737	0.039	0.000	1.064	0.013	0.000
0.7	0.923	0.083	0.000	1.547	0.033	0.044
Heavy Censoring($\lambda_c = 0.02$)						
0	0.757	0.049	0.000	0.967	0.020	0.005
0.2	0.801	0.057	0.000	1.040	0.022	0.003
0.5	0.948	0.073	0.000	1.349	0.032	0.013
0.7	1.133	0.090	0.002	1.889	0.036	0.096

Model Error: $[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^T * (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})]^{0.5}, \boldsymbol{\beta^*} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$
True $\boldsymbol{\beta^*}$: $\boldsymbol{\beta^*} = (-1, 1, -1.5, 2, 1.75, -1.25, 0, 0, ..., 0)^T$
FPR: False positive rate = (# of false positive covariates)/(total # of covariates)
FNR: False negative rate = (# of false negative covariates)/(total # of covariates)
Corr.: Correlation

Table 6.2: Simulation results when $\alpha^* = 0.9$ and $\theta^* = 6$

	$n = 200, p = 100$			$n = 200, p = 300$		
Corr. ρ	Model Error	FPR	FNR	Model Error	FPR	FNR
Mild Censoring($\lambda_c = 0.0002$)						
0	0.641	0.044	0.000	0.811	0.015	0.000
0.2	0.707	0.052	0.000	0.900	0.019	0.000
0.5	0.891	0.075	0.000	1.208	0.028	0.000
0.7	1.093	0.096	0.000	1.900	0.035	0.103
Moderate Censoring($\lambda_c = 0.002$)						
0	0.701	0.052	0.000	0.856	0.016	0.000
0.2	0.750	0.056	0.000	0.953	0.020	0.000
0.5	0.914	0.076	0.000	1.300	0.031	0.003
0.7	1.142	0.099	0.001	1.989	0.033	0.129
Heavy Censoring($\lambda_c = 0.02$)						
0	0.886	0.063	0.000	1.102	0.022	0.003
0.2	0.948	0.067	0.000	1.256	0.029	0.007
0.5	1.096	0.087	0.000	1.702	0.037	0.044
0.7	1.355	0.106	0.005	2.318	0.032	0.238

Model Error: $[(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})^T * (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)]^{0.5}, \boldsymbol{\beta}^* = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$
True $\boldsymbol{\beta}^*$: $\boldsymbol{\beta}^* = (-1, 1, -1.5, 2, 1.75, -1.25, 0, 0, ..., 0)^T$
FPR: False positive rate = (# of false positive covariates)/(total # of covariates)
FNR: False negative rate = (# of false negative covariates)/(total # of covariates)
Corr.: Correlation

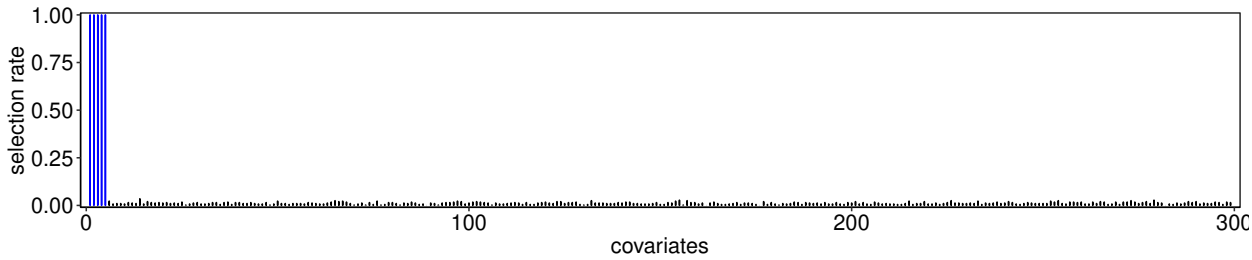
correlation is high (i.e., $\rho = 0.7$), the rates of correct selection are between about 0.75 and 1. Furthermore, it is interesting to note that true positives x_3 and x_4 have higher chances of being correctly selected in this multicollinearity setting. There are two reasons: (1)The negative effects ($\beta_2 = -1.5$) will be diminished by the neighboring positive effects ($\beta_1 = 1$ and $\beta_3 = 2$) as high correlation exists between the neighboring covariates. Thus the chance to be selected will be decreased, which is also true for β_5. (2)As each covariate is generated with the same variance, then the magnitude of the coefficient will represent the variable importance. Larger coefficients tend to have higher chances to be selected under the same penalization. However, this is not necessarily true under the multicollinearity setting as the magnitude of coefficient might be diminished by highly correlated variables.

In our study, as a random variable is adopted to represent the uncure fraction, we conjecture that this representation will bring robustness for the model estimation against any link function for the uncure fraction. As an empirical evidence, we generate data by different true link functions for the uncure fraction and the simulation results given in Table 6.3 show that false positive rates and false negative rates are small regardless of the form the uncure fraction.

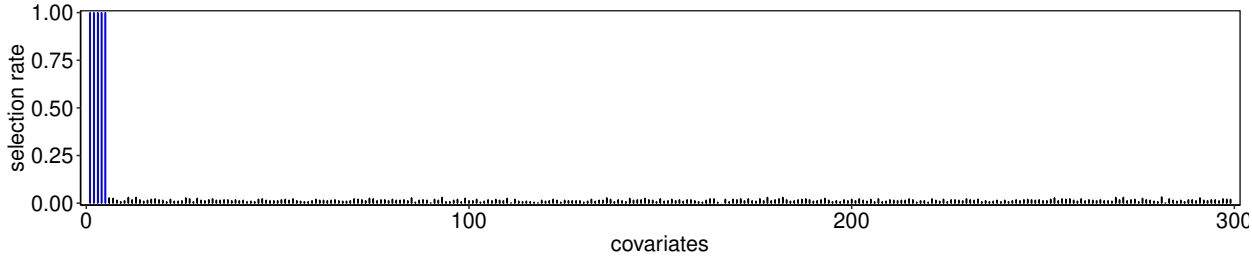
6.4.2 Application to TCGA breast cancer data

The TCGA breast cancer dataset contain 607 patients with about $40,000$ microarray expressions (i.e., covariates). Figure 6.3 is the survival curve for the breast cancer patient data. As seen from the graph, survival curve levels off around the rate about 0.6 after about 120 months follow-up. This phenomenon suggests the existence of long-term survivors in the data, meaning that a marginal cure rate model is appropriate to use.

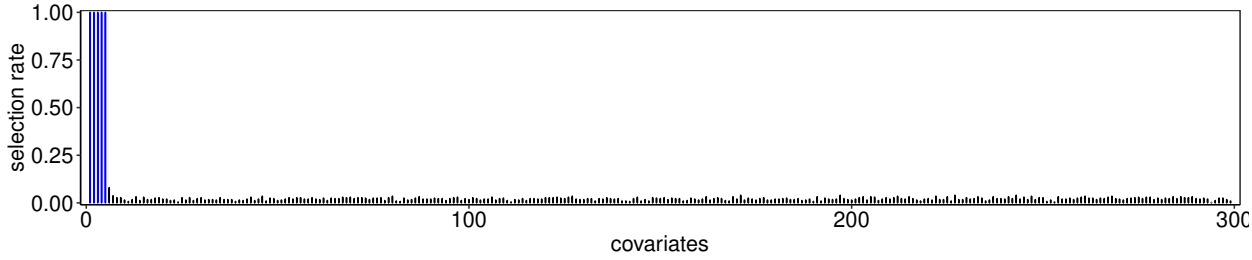
We preselect 8,000 microarrays by Cox's score (Witten and Tibshirani, 2010) for the purpose of analysis. The average correlation for all pairs of those 8000 microarrays is 0.05 with maximum correlation is 0.99 and minimum is -0.97. The histogram of these correlations is given in Figure 6.4. We can observe that about 75% of these correlations are falling in $[-0.06, 0.19]$.



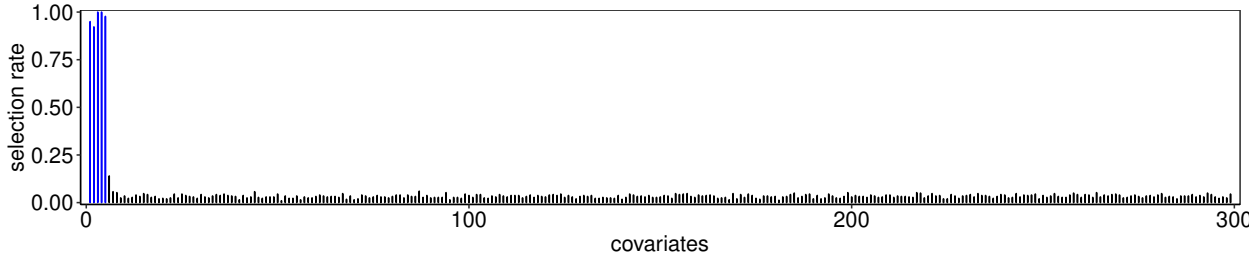
(a) $\rho = 0$



(b) $\rho = 0.2$

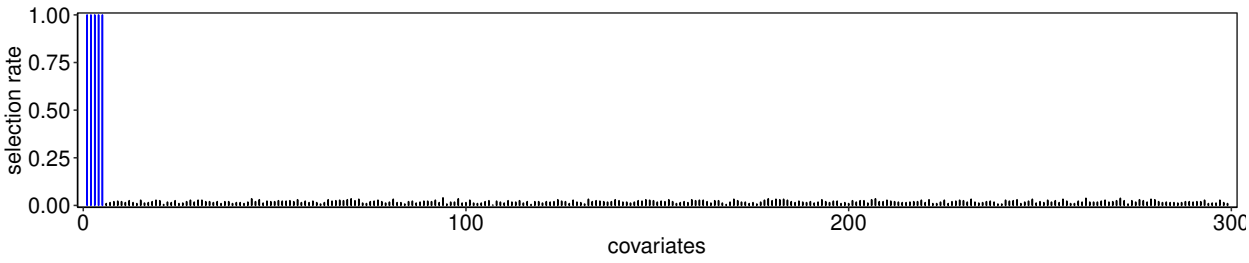


(c) $\rho = 0.5$

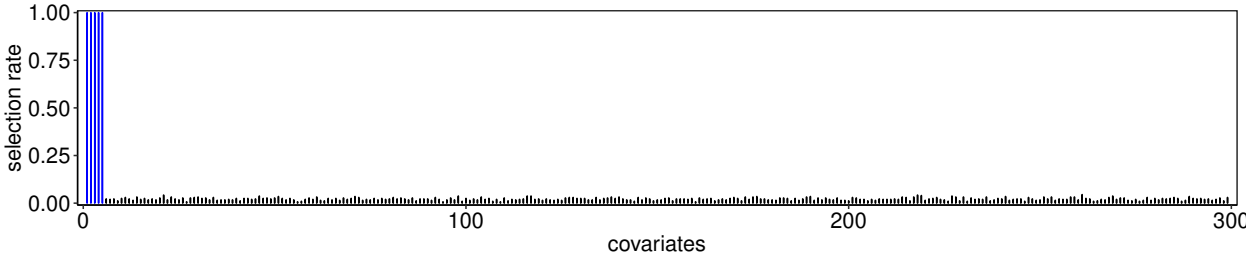


(d) $\rho = 0.7$

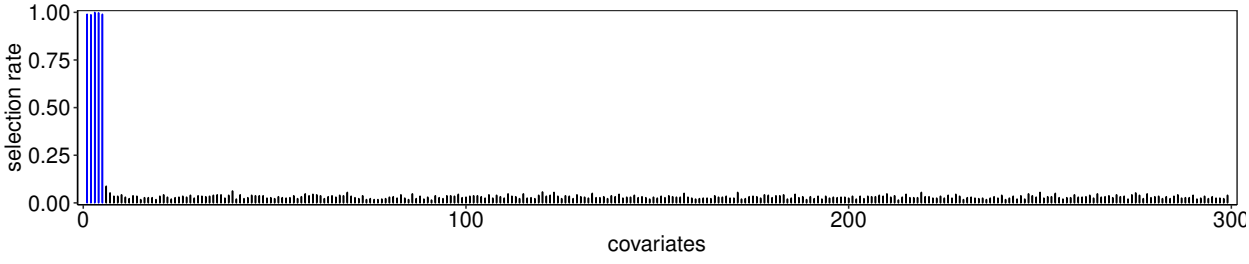
Figure 6.1: The rate of each covariate selected by the model for 500 replicates with $\alpha^* = 1.1, \theta^* = 6$ and $\lambda_c = 0.0002$, where only the first 5 covariates are used for data generation



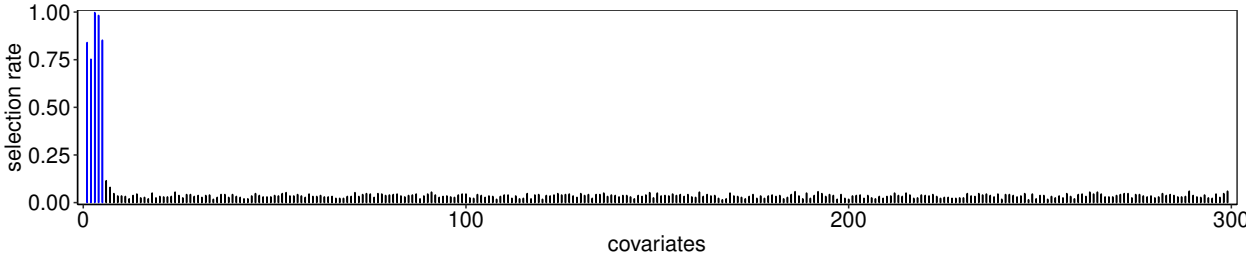
(a) $\rho = 0$



(b) $\rho = 0.2$



(c) $\rho = 0.5$



(d) $\rho = 0.7$

Figure 6.2: The rate of each covariate selected by the model for 500 replicates with $\alpha^* = 1.1, \theta^* = 6$ and $\lambda_c = 0.02$, where only the first 5 covariates are used for data generation

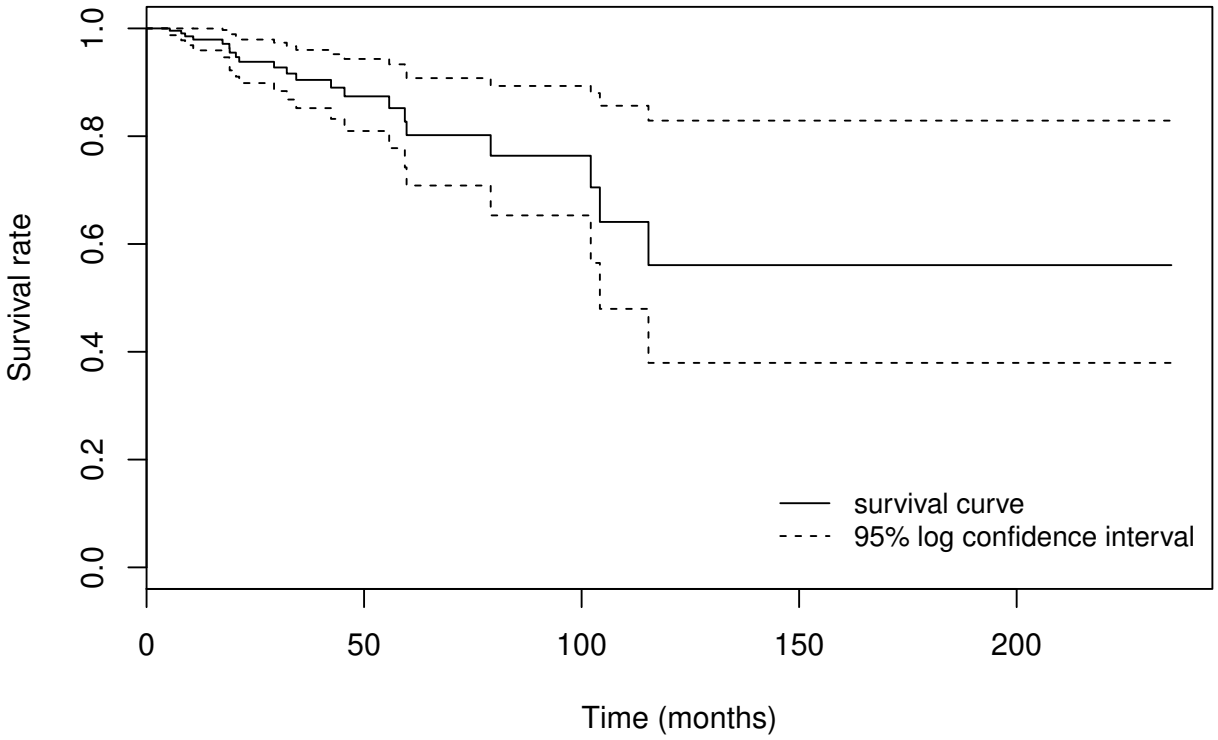


Figure 6.3: Survival curve for the breast cancer patients for TCGA data

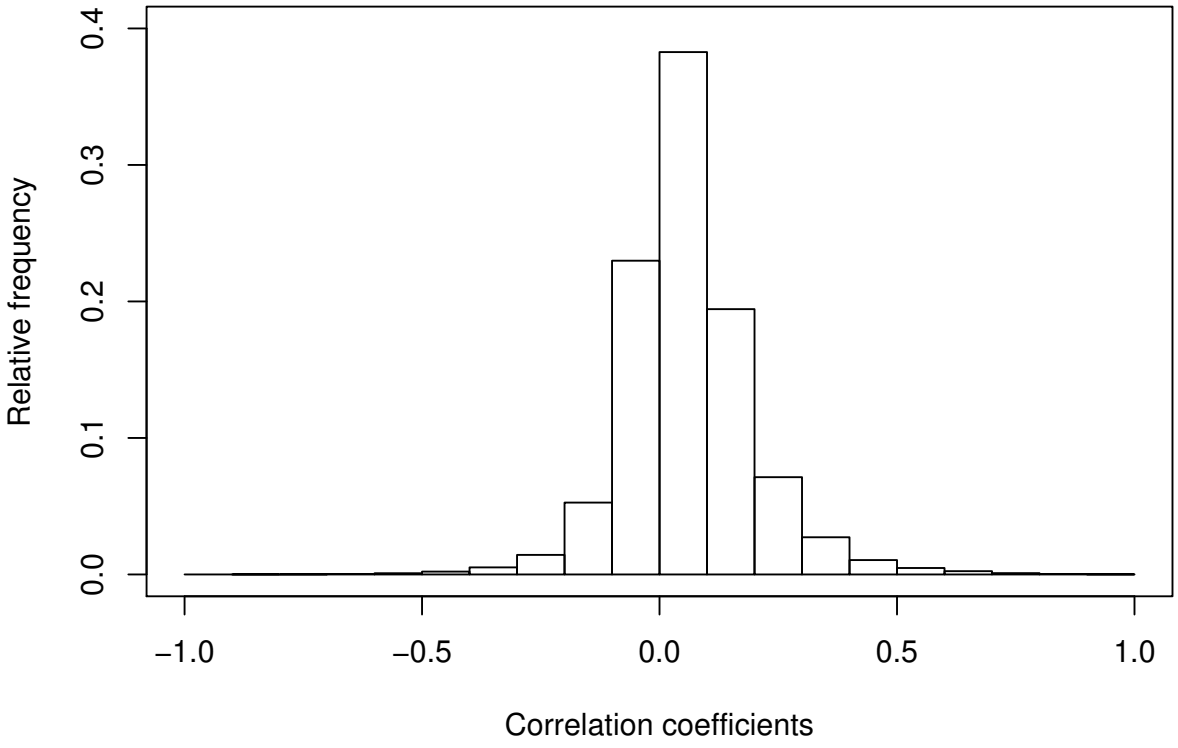


Figure 6.4: The histogram of all Pearson's ρ for all $\binom{8000}{2}$ paired microarrays

Table 6.3: Robustness of the proposed method with various π_i^*, $\alpha^* = 1.1$,$n = 200$,$p = 300$ and $\rho = 0$

	Working uncure fraction					
	$\pi^{-1.1} \sim Exp(\theta, 1)$			$\pi_i = constant$		
True link function	Model Error	FPR	FNR	Model Error	FPR	FNR
$\pi_i^* = 0.75$	1.123	0.026	0.022	1.174	0.028	0.034
$\pi_i^* = 1 - \exp\big\{-\exp\{3.5 - 2x_1 - 1.5x_2\}\big\}$	0.976	0.021	0.002	1.006	0.022	0.008
$\pi_i^* = \exp\big\{-\exp\{-4 + 2x_1 + 1.5x_2\}\big\}$	1.094	0.026	0.009	1.152	0.028	0.020
$\pi_i^* = \Phi(3.5 - 2x_1 - 1.5x_2)$	1.096	0.024	0.009	1.096	0.026	0.016
$\text{logit}(\pi_i) = 3.5 - 2x_1 - 1.5x_2$	1.174	0.028	0.023	1.182	0.030	0.040
$\pi^{-1.1} \sim Exp(\theta^*, 1)$	0.964	0.020	0.005	0.979	0.020	0.006

Model Error: $[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})^T * (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta^*})]^{0.5}, \boldsymbol{\beta^*} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$
True $\boldsymbol{\beta}$: $\boldsymbol{\beta^*} = (-1, 1, -1.5, 2, 1.75, -1.25, 0, 0, ..., 0)^T$
FPR: False positive rate = (# of false positive covariates)/(total # of covariates)
FNR: False negative rate = (# of false negative covariates)/(total # of covariates)
Φ: The CDF of standard normal distribution

Table 6.4: Variable selection results with estimated coefficients

Microarray	Parameter
cg02421553	−3.085
cg08150755	−0.035
cg14059988	−0.037
cg27047459	−0.015
cg22191803	−0.010
cg14458509	−0.025

As most of the correlations are small, our proposed model is expected to be reliable. The real data analysis result is given in Table 6.4. It suggests that there are 6 important microarrays related to the overall hazard rate of breast cancer patients with negative coefficients. Among those 6 microarrays, *cg02421553* is the one with the most great effect size, and all of the 6 are with negative coefficients.

6.5 Discussion

A few studies in the literature focus on cure rate model with high-dimensional data. This study is the first one to propose a feasible method to handle and select important covariates that are related to the marignal mean hazard rate from the high-dimensional data. A quadrature programming algorithm proposed by Tibshirani (2009), orginally used for high-dimensional Cox proportional hazard model, is extended to our model.

Chapter 7

Conclusion

Existing literature has overlooked the explanation of covariate effects on the overall survival outcome when long-term survivors are present in the data. The classical cure rate models offer two sets of parameters, which could only make interpretations on the probability of being cured and on the hazard rate of the uncured subpopulation. Therefore, the way to interpret the covariate effects under classical cure rate model is pretty restricted. It becomes very challenging for studies that are attempting to directly interpret the covariate effects on the marginal survival rate or hazard rate.

Based on the classical cure rate model, we develop a new type of model, marginal cure rate models with novel parameterizations that can relate covariates directly to the marginal mean survival rate or marginal mean hazard rate. The proposed models yield a nice solution to interpret the covariate effects on the marginal survival outcomes. Our models based on these novel parameterizations together with the Weibull assumption of baseline hazard can be easily fitted by using routine statistical software such as SAS NLMIXED procedure with a minimum programming effort.

To relax the restriction of Weibull baseline hazard assumption for the uncured subpopulation, we extend the model and propose a semi-parametric technique using Berstein polynomials. This spline method enjoys the benefits of fast implementation as well as no need to decide the spline knots. After comparing with the results of Weibull method, we

find that the proposed semi-parametric method performs well in the simulation study and the liver cancer application. For the future study, we need to strengthen the theoretic properties of sieve likelihood method, such as convergence rate. Moreover, intercept term of the marginal mean hazard parameters is not identifiable in this semi-parametric model. Even though intercept term is usually not the of interest to interpret, it is also deserved to be further studied.

Lastly, motivated by the microarray data of breast cancer patients from The Cancer Genome Atlas (TCGA), we further extend the marginal mean hazard rate model to high-dimensional data in the sense of a massive number of covariates (i.e. genes). It is worth to note that current work is focused on the lasso penalty for variable selection. This penalty is not appealing when covariates are related or having groupwise structure. Future work could extend the current marginal mean hazard model to the elastic net or other penalties for much more flexibility.

Bibliography

Ross L Prentice and John D Kalbfleisch. Hazard rate models with covariates. *Biometrics*, pages 25–39, 1979.

Richard Sposto. Cure model analysis in cancer: an application to data from the children's cancer group. *Statistics in medicine*, 21(2):293–312, 2002.

Theodora Bejan-Angoulvant, Anne-Marie Bouvier, Nadine Bossard, Aurelien Belot, Valérie Jooste, Guy Launoy, and Laurent Remontet. Hazard regression model and cure rate model in colon cancer relative survival trends: are they telling the same story? *European journal of epidemiology*, 23(4):251–259, 2008.

Alessando Cucchetti, Alessandro Ferrero, Matteo Cescon, Matteo Donadon, Nadia Russo-lillo, Giorgio Ercolani, Giacomo Stacchini, Federico Mazzotti, Guido Torzilli, and Antonio Daniele Pinna. Cure model survival analysis after hepatic resection for colorectal liver metastases. *Annals of surgical oncology*, 22(6):1908–1914, 2015.

John W Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53, 1949.

Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

Vern T Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

Vernon T Farewell. Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262, 1986.

Jeremy MG Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, pages 899–907, 1995.

Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

A Azzalini. Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, 81(4):767–775, 1994.

Patrick J Heagerty, Scott L Zeger, et al. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.

Jeffrey M Albert, Wei Wang, and Suchitra Nelson. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical methods in medical research*, 23(3):257–278, 2014.

D Leann Long, John S Preisser, Amy H Herring, and Carol E Golin. A marginalized zero-inflated poisson regression model with overall exposure effects. *Statistics in medicine*, 33 (29):5151–5165, 2014.

D Leann Long, John S Preisser, Amy H Herring, and Carol E Golin. A marginalized zero-inflated poisson regression model with random effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(5):815–830, 2015.

Valerie A Smith, John S Preisser, Brian Neelon, and Matthew L Maciejewski. A marginalized two-part model for semicontinuous data. *Statistics in medicine*, 33(28):4891–4903, 2014.

David Todem, KyungMann Kim, and Wei-Wen Hsu. Marginal mean models for zero-inflated count data. *Biometrics*, 72(3):986–994, 2016.

Wei Wang and Michael E Griswold. Estimating overall exposure effects for the clustered and censored outcome using random effect tobit regression models. *Statistics in medicine*, 35 (27):4948–4960, 2016.

Wei Wang and Michael E Griswold. Natural interpretations in tobit regression models using marginal estimation methods. *Statistical methods in medical research*, 26(6):2622–2632, 2017.

Xinyan Fan, Mengque Liu, Kuangnan Fang, Yuan Huang, and Shuangge Ma. Promoting structural effects of covariates in the cure rate model with penalization. *Statistical methods in medical research*, 26(5):2078–2092, 2017.

David R Cox. Models and life-tables regression. *JR Stat. Soc. Ser. B*, 34:187–220, 1972.

Lu Wang, Pang Du, and Hua Liang. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68(3):726–735, 2012.

Wei-Wen Hsu, David Todem, and KyungMann Kim. A sup-score test for the cure fraction in mixture models for long-term survivors. *Biometrics*, 72(4):1348–1357, 2016.

Yun Zhao, Andy H Lee, Kelvin KW Yau, Valerie Burke, and Geoffrey J McLachlan. A score test for assessing the cured proportion in the long-term survivor mixture model. *Statistics in medicine*, 28(27):3454–3466, 2009.

Yingwei Peng, Keith BG Dear, and JW Denham. A generalized f mixture model for cure rate estimation. *Statistics in medicine*, 17(8):813–830, 1998.

Chin-Shang Li and Jeremy MG Taylor. A semi-parametric accelerated failure time cure model. *Statistics in medicine*, 21(21):3235–3247, 2002.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Odd O Aalen, Håkon K Gjessing, et al. Understanding the shape of the hazard rate: A process point of view (with comments and a rejoinder by the authors). *Statistical Science*, 16(1):1–22, 2001.

Toji Makino. Mean hazard rate and its application to the normal approximation of the weibull distribution. *Naval research logistics quarterly*, 31(1):1–8, 1984.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2006.

Robert Gentleman and Charles J Geyer. Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81(3):618–623, 1994.

Ross A Maller and S Zhou. Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association*, 89(428):1499–1506, 1994.

Anthony YC Kuk and Chen-Hsin Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541, 1992.

Yingwei Peng and Keith BG Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

Hao Liu and Yu Shen. A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association*, 104(487):1168–1178, 2009.

Qingning Zhou, Tao Hu, and Jianguo Sun. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, 112(518):664–672, 2017.

Jesús M Carnicer and Juan Manuel Pena. Shape preserving representations and optimality of the bernstein basis. *Advances in Computational Mathematics*, 1(2):173–196, 1993.

Jorge Delgado and Juan Manuel Pena. Optimality of bernstein representations for computational purposes. *Reliable Computing*, 17(1):1–10, 2012.

Muhtarjan Osman and Sujit K Ghosh. Nonparametric regression models for right-censored data using bernstein polynomials. *Computational Statistics & Data Analysis*, 56(3):559–573, 2012.

Jian Huang and AJ Rossini. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, 92(439): 960–967, 1997.

DR Anderson and K Burnham. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, page 63, 2004.

Andreas Raue, Johan Karlsson, Maria Pia Saccomani, Mats Jirstrand, and Jens Timmer. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*, 30(10):1440–1448, 2014.

Daniela M Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51, 2010.

James O Ramsay et al. Monotone regression splines in action. *Statistical science*, 3(4): 425–441, 1988.

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Robert J Tibshirani. Univariate shrinkage in the cox model for high dimensional data. *Statistical applications in genetics and molecular biology*, 8(1):1–18, 2009.

Jiang Gui and Hongzhe Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005.

Xiang Liu, Yingwei Peng, Dongsheng Tu, and Hua Liang. Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in medicine*, 31(24):2882–2891, 2012.

Abdullah Masud, Wanzhu Tu, and Zhangsheng Yu. Variable selection for mixture and promotion time cure rate models. *Statistical methods in medical research*, 27(7):2185–2199, 2018.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Hansheng Wang and Chenlei Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

Gyuhyeong Goh and Dipak K Dey. Asymptotic properties of marginal least-square estimator for ultrahigh-dimensional linear regression models with correlated errors. *The American Statistician*, 73(1):4–9, 2019.

Appendix A

SAS Code

A.1 Marginal mean survival rate model

```
%let N = 200;
%let MCsamp = 1000;
%let r0=-1.5;
%let r1=1;
%let r2=4;
%let b1=-0.5;
%let b2=-2;
%let alpha=1.1;
%let lambda=0.01;
data simulation (drop=i censor1 censor2);
do sampleID = 1 to &MCsamp;
do i= 1 to &N;
ss=rand('uniform');
x1=rand('normal');
x2=rand('unif');
p=exp(&r0+&r1*x1+&r2*x2)/(0.5+exp(&r0+&r1*x1+&r2*x2));
```

```
conmean=exp(&b1*x1+&b2*x2);
tt=(-(log(1-ss))/(&lambda*conmean))**(1/&alpha);
g=rand('binom',p,1);
censor1 = rand('EXPONENTIAL');
censor2 = censor1/0.001;
censor = min(800,censor2);
curet=100000;
if g=1 then truet=tt; else truet=curet;
if truet le censor then d=1; else d=0;
if d=1 then obt=truet; else obt=censor;
output;
end;
end;
run;
%macro ODSOff();
ods graphics off;
ods exclude all;
ods results off;
options nonotes;
%mend;
ods graphics on;
ods exclude none;
ods results;
%mend;
%ODSOff
proc nlmixed data=simulation gconv=1e-20 fconv=1e-20;
by SampleID;
parms r0=-1.5 r1=1 r2=4 b1=-0.5 b2=-2 alpha=1.1 lambda=0.01;
bounds alpha>0, lambda>0;
```

```
eta1=r0+r1*x1+r2*x2;
eta2=b1*x1+b2*x2;
p=exp(eta1)/(0.5+exp(eta1));
t=obt;
s0=exp(-lambda*(t**alpha));
if d=1 then
loglik=log(p)+log(lambda)+log(alpha)+log(s0)*exp(eta2)+(alpha-1)*log(t)+eta2;
else loglik=log(1-p+p*(s0**exp(eta2)));
model t general(loglik);
ods output ParameterEstimates=estimate;
run;
%ODSOn
proc sql;
create table want as
select *
from estimate
group by sampleID
having min(abs(estimate)) ge 0;
quit;
data want2;
set want;
if parameter='alpha' then true=&alpha;
if parameter='lambda' then true=&lambda;
if parameter='b1' then true=&b1;
if parameter='b2' then true=&b2;
if parameter='r0' then true=&r0;
if parameter='r1' then true=&r1;
if parameter='r2' then true=&r2;
if true ≤ Lower or true ≥ Upper then inconfi=0; else inconfi=1;
```

```
bias=abs((estimate-true)/true);
run;
proc summary data=want2 print;
var estimate;
run;
PROC MEANS DATA=want2 MEAN STD MAXDEC=5;
CLASS parameter;
VAR estimate StandardError bias inconfi;
run;
```

A.2 Marginal mean hazard rate model

```
%let N = 800;
%let MCsamp = 1000;
%let b0=-4;
%let b1=1;
%let b2=-2;
%let r0=1;
%let r1=2;
%let r2=1.5;
%let alpha=0.75;
%let lambda=0.1;
data simulation (drop=i censor1 censor2);
do sampleID = 1 to &MCsamp;
do i= 1 to &N;
ss=rand('uniform');
x1=rand('bernoulli',0.5);
x2=rand('normal");
p=exp(&r0+&r1*x1+&r2*x2)/(1+exp(&r0+&r1*x1+&r2*x2));
```

```
marmeanh=exp(&b0+&b1*x1+&b2*x2);
vv=gamma(2-1/&alpha);
condmean=1/&lambda*(marmeanh/(&alpha*p*gamma(2-1/&alpha)))**(&alpha);
tt=(-(log(1-ss))/(&lambda*condmean))**(1/&alpha);
g=rand('binom',p,1);
censor1 = rand('EXPONENTIAL');
censor2 = censor1/0.0002;
censor = min(1000,censor2);
curet=100000;
if g=1 then truet=tt; else truet=curet;
if truet le censor then d=1; else d=0;
if d=1 then obt=truet; else obt=censor;
output;
end;
end;
run;
%macro ODSOff();
ods graphics off;
ods exclude all;
ods results off;
options nonotes;
%mend;
%macro ODSOn();
ods graphics on;
ods exclude none;
ods results;
%mend;
%ODSOff
proc nlmixed data=simulation gconv=1e-20 fconv=1e-20 tech=QUANEW;
```

```
by SampleID;
parms b0=-5 b1=1 b2=-2 r0=1 r1=1.5 r2=2 alpha=1.1;
bounds alpha>0;
eta1=b0+b1*x1+b2*x2;
eta2=r0+r1*x1+r2*x2;
p=exp(eta2)/(1+exp(eta2));
mm=exp(eta1);
lameu=(mm/(alpha*p*gamma(2-1/alpha)))**(alpha);
t=obt;
su=exp(-lameu*(t**alpha));
if d=1 then loglik=log(p)+log(lameu)+log(alpha)-(t**alpha)*lameu+(alpha-1)*log(t);
else loglik=log(1-p+p*su);
model t general(loglik);
ods output ParameterEstimates=estimate;
run;
%ODSOn
proc sql;
create table want as
select *
from estimate
group by sampleID
having min(abs(estimate)) ge 0;
quit;
data want2;
set want;
if parameter='alpha' then true=&alpha;
if parameter='lambda' then true=&lambda;
if parameter='b0' then true=&b0;
if parameter='b1' then true=&b1;
```

if parameter='b2' then true=&b2;

if parameter='r0' then true=&r0;

if parameter='r1' then true=&r1;

if parameter='r2' then true=&r2;

if true $\leq$ Lower or true $\geq$ Upper then inconfi=0; else inconfi=1;

mse=(estimate-true)**2;

bias=abs((estimate-true)/true);

run;

PROC MEANS DATA=want2 MEAN STD MAXDEC=5 ;

CLASS parameter;

VAR mse estimate StandardError bias inconfi;

run;

A.3 True censoring rate for simulations

A.3.1 Marginal mean survival rate model

For marginal mean survival rate model, the hazard function for the uncured subpopulation is assumed to be

$$\begin{aligned} h_u(t_i) &= h_{u0}(t_i) \exp\left\{\boldsymbol{\eta'} \boldsymbol{x_i}\right\} \\ &= 1.1 \times 0.01 \times t_i^{0.1} \times \exp\{-0.5X_{i1} - 2X_{i2}\} \end{aligned}$$

then survival function for uncured subpopulation is

$$\begin{aligned} S_u(t_i) &= S_{u0}(t_i)^{\exp\{\boldsymbol{\eta'} \boldsymbol{x_i}\}} \\ &= e^{-0.01 \times t_i^{1.1} \times \exp\{-0.5X_{i1} - 2X_{i2}\}} \end{aligned}$$

The censoring time t_c is generated from exponential distribution with $f(t_c) = \lambda_c e^{-\lambda_c t_c}$.

As t_c is independent of t_i, then $f(t_i, t_c) = f_u(t_i)f(t_c)$. The true censoring rate CR is

$$\begin{aligned} CR_i &= P(t_i \geq t_c | X_i) \\ &= \int_0^\infty \int_{t_c}^\infty f(t_i, t_c) dt_i dt_c \\ &= \int_0^\infty \int_{t_c}^\infty f_u(t_i) f(t_c) dt_i dt_c \\ &= \int_0^\infty f(t_c) S_u(t_c) dt_c \\ &= \int_0^\infty \lambda_c e^{-\lambda_c t_c - 0.01 \times t_c^{\{1.1\}} \times \exp\{-0.5X_{i1} - 2X_{i2}\}} dt_c \end{aligned}$$

where $\lambda_c = 0.002, 0.001$ and 0.0002 respectively. There is no closed form for above integration. We could use numerical method to approximate the integration for each observation i and then get the true average censoring rate. By calculation, $CR = 32.88\%, 21.48\%$ and 9.18% for the uncured subpopulation respectively.

A.3.2 Marginal mean hazard rate model

When data is generated from the standard cure rate model, the hazard function for the uncured subpopulation is assumed to be

$$\begin{aligned} h_u(t_i) &= h_{u0}(t_i) \exp\left\{\boldsymbol{\eta' x_i}\right\} \\ &= 0.75 \times 0.1 \times t_i^{-0.25} \times \exp\{-4 + X_{i1} - 2X_{i2}\} \end{aligned}$$

then survival function for uncured subpopulation is

$$\begin{aligned} S_u(t_i) &= S_{u0}(t_i)^{\exp\{\boldsymbol{\eta' x_i}\}} \\ &= e^{-0.1 \times t_i^{0.75} \times \exp\{-4 + X_{i1} - 2X_{i2}\}} \end{aligned}$$

We also assume the censoring time t_c follows exponential distribution, then $f(t_c) = \lambda_c e^{-\lambda_c t_c}$. As t_c is independent of t_i, then $f(t_i, t_c) = f_u(t_i)f(t_c)$. The true censoring rate CR is

$$
\begin{aligned}
CR_i &= P(t_i \geq t_c | X_i) \\
&= \int_0^\infty \int_{t_c}^\infty f(t_i, t_c) dt_i dt_c \\
&= \int_0^\infty \int_{t_c}^\infty f_u(t_i) f(t_c) dt_i dt_c \\
&= \int_0^\infty f(t_c) S_u(t_c) dt_c \\
&= \int_0^\infty \lambda_c e^{-\lambda_c t_c - 0.1 \times t_c^{\{0.75\}} \times \exp\{-4 + X_{i1} - 2X_{i2}\}} dt_c
\end{aligned}
$$

where $\lambda_c = 0.002, 0.001$ and 0.0002 respectively. There is no closed form for above integration. We could use numerical method to approximate the integration for each observation i and then get the true average censoring rate. By calculation, $CR = 14.69\%$, 10.18% and 8.16% respectively.

When data is generated from the marginal cure rate model, hazard function for the uncured subpopulation is

$$
\begin{aligned}
h_u(t_i) &= \alpha t_i^{\alpha-1} \Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha \pi_i \Gamma(2 - \frac{1}{\alpha})} \Big]^\alpha \\
&= 0.75 \times t_i^{-0.25} \Big[\frac{\exp\{-4 + X_{i1} - 2X_{i2}\}}{0.75 \times \pi_i \times \Gamma(2/3)} \Big]^{0.75},
\end{aligned}
$$

then survival function for uncured subpopulation is

$$
\begin{aligned}
S_u(t_i) &= S_{u0}(t_i)^{\exp\{\boldsymbol{\eta' x_i}\}} \\
&= \exp\Big\{ - t_i^{0.75} \times \Big[\frac{\exp\{-4 + X_{i1} - 2X_{i2}\}}{0.75 \times \pi_i \times \Gamma(2/3)} \Big]^{0.75} \Big\}
\end{aligned}
$$

then the true censoring rate CR_i given covariates X_i is

$$\begin{aligned} CR_i &= P(t_i \geq t_c | X_i) \\ &= \int_0^\infty \int_{t_c}^\infty f(t_i, t_c) dt_i dt_c \\ &= \int_0^\infty \int_{t_c}^\infty f_u(t_i) f(t_c) dt_i dt_c \\ &= \int_0^\infty f(t_c) S_u(t_c) dt_c \\ &= \int_0^\infty \lambda_c \exp\Big\{ -\lambda_c t_c - t_i^{0.75} \times \Big[\frac{\exp\{-4 + X_{i1} - 2X_{i2}\}}{0.75 \times \pi_i \times \Gamma(2/3)} \Big]^{0.75} \Big\} dt_c \end{aligned}$$

The above calculation can also be calculated by using numerical method for $\lambda_c = 0.002, 0.001$ and 0.0002. By calculation, $CR = 18.06\%, 13.01\%$ and 7.67% for the uncured subpopulation respectively.

A.3.3 Semi-parametric marginal mean hazard rate model

When data is generated from the marginal cure rate model, hazard function for the uncured subpopulation is

$$\begin{aligned} h_u(t_i) &= \alpha t_i^{\alpha-1} \Big[\frac{e^{\boldsymbol{\beta' x_i}}}{\alpha \pi_i \Gamma(2 - \frac{1}{\alpha})} \Big]^\alpha \\ &= 0.8 \times t_i^{-0.2} \Big[\frac{\exp\{-1.5 - 1.25X_{i1} - 0.75X_{i2}\}}{0.8 \times \pi_i \times \Gamma(3/4)} \Big]^{0.8}, \end{aligned}$$

then survival function for uncured subpopulation is

$$\begin{aligned} S_u(t_i) &= S_{u0}(t_i)^{\exp\{\boldsymbol{\eta' x_i}\}} \\ &= \exp\Big\{ -t_i^{0.8} \times \Big[\frac{\exp\{-1.5 - 1.25X_{i1} - 0.75X_{i2}\}}{0.8 \times \pi_i \times \Gamma(3/4)} \Big]^{0.8} \Big\} \end{aligned}$$

then the true censoring rate CR_i given covariates X_i is

$$\begin{aligned}
CR_i &= P(t_i \geq t_c | X_i) \\
&= \int_0^\infty \int_{t_c}^\infty f(t_i, t_c) dt_i dt_c \\
&= \int_0^\infty \int_{t_c}^\infty f_u(t_i) f(t_c) dt_i dt_c \\
&= \int_0^\infty f(t_c) S_u(t_c) dt_c \\
&= \int_0^\infty \lambda_c \exp\Big\{ -\lambda_c t_c - t_i^{0.8} \times \Big[\frac{\exp\{-1.5 - 1.25X_{i1} - 0.75X_{i2}\}}{0.8 \times \pi_i \times \Gamma(3/4)}\Big]^{0.8} \Big\} dt_c
\end{aligned}$$

The above calculation can also be calculated by using numerical method for $\lambda_c = 0.1, 0.01$ and 0.001. By calculation, the $CR = 29.80\%, 7.62\%$ and 4.01% for the uncured subpopulation respectively.

A.4 EM algorithm for marginal cure rate model

Suppose we have independently observed survival data $\{t_i, y_i, \delta_i, \boldsymbol{z_i}, \boldsymbol{w_i}\}$, where t_i is the survival time for individual i. The y_i is the cured indicator, $y_i = 1$ if cured and 0 otherwise. The δ_i is the censoring indicator ($\delta_i = 1$, noncensored; $\delta_i = 0$, censored). The covariates $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$ are related to the uncure rate π_i and the hazard rate $h_u(t_i)$, respectively. Then the completely likelihood function for the marginal cure rate model is

$$\mathcal{L}_c(\alpha, \lambda, \boldsymbol{\eta}, \boldsymbol{\gamma} | \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{W}, \boldsymbol{Z}) = \prod_{i=1}^n \Big[\pi_i(\boldsymbol{w_i}) f_u(t_i | \boldsymbol{z_i})\Big]^{\delta_i(1-y_i)} \Big\{ [1 - \pi_i(\boldsymbol{w_i})]^{y_i} [\pi_i(\boldsymbol{w_i}) S_u(t_i | \boldsymbol{z_i})]^{1-y_i} \Big\}^{1-\delta_i}$$

Taking log function on both sides, we could get the completely log likelihood function as

follows,

$$\begin{aligned}\ell_c = \sum_{i=1}^{n} \Big\{ &\big[\delta_i(1-y_i) + (1-\delta_i)(1-y_i)\big]\pi_i(\boldsymbol{w_i}) + (1-\delta_i)y_i \log\big[1-\pi_i(\boldsymbol{w_i})\big] \\ &+ \delta_i(1-y_i)\log\big[h_u(t_i|\boldsymbol{z_i})S_u(t_i|\boldsymbol{z_i})\big] + (1-\delta_i)(1-y_i)\log S_u(t_i|\boldsymbol{z_i})\Big\} \\ = \sum_{i=1}^{n} &(1-y_i)\pi_i(\boldsymbol{w_i}) + y_i \log\big[1-\pi_i(\boldsymbol{w_i})\big] + \delta_i \log h_u(t_i|\boldsymbol{z_i}) + (1-y_i)\log S_u(t_i|\boldsymbol{z_i})\end{aligned} \tag{A.1}$$

The second equality holds in Equation A.1 as (i)if $y_i = 1$ then $\delta_i = 0$, (ii)the set of values of y_i and δ_i is $\{0, 1\}$. Applying the EM algorithm, the E step calculate the expectation of $1 - y_i$ given current estimates of $(\alpha, \lambda, \boldsymbol{\eta}, \boldsymbol{\gamma})$. Let $g_i = E\big[(1-y_i)|\alpha, \lambda, \boldsymbol{\eta}, \boldsymbol{\gamma}\big]$, then $g_i = P(y_i = 0)$. It's obvious that $g_i = 1$ if $\delta_i = 1$. For the case $\delta_i = 0$,

$$\begin{aligned} g_i &= P(y_i = 0|\delta_i = 0) \\ &= P(y_i = 0|T_i > t_i) \\ &= 1 - P(y_i = 1|T_i > t_i) \\ &= 1 - \frac{P(T_i > t_i|y_i = 1)P(y_i = 1)}{P(T_i > t_i)} \\ &= 1 - \frac{P(y_i = 1)}{P(T_i > t_i)} \\ &= \frac{\pi_i(\boldsymbol{w_i})S_u(t_i|\boldsymbol{z_i})}{1 - \pi_i(\boldsymbol{w_i}) + \pi_i(\boldsymbol{w_i})S_u(t_i|\boldsymbol{z_i})} \end{aligned}$$

where T_i is the actual survival time of individual i. Combining the cases of $\delta_i = 1$ and $\delta_i = 0$, then we could rewrite

$$g_i = \delta_i + \frac{(1-\delta_i)\pi_i(\boldsymbol{w_i})S_u(t_i|\boldsymbol{z_i})}{1 - \pi_i(\boldsymbol{w_i}) + \pi_i(\boldsymbol{w_i})S_u(t_i|\boldsymbol{z_i})} \tag{A.2}$$

M step maximize the expected completely likelihood given current estimates, which could be written as follows according to Equation A.1 and A.2,

$$\ell_c = \sum_{i=1}^{n} \Big[g_i \log \pi_i(\boldsymbol{w_i}) + (1-g_i)\log(1-\pi_i(\boldsymbol{w_i}))\Big] + \sum_{i=1}^{n} \Big[g_i S_u(t_i|\boldsymbol{z_i}) + \delta_i \log h_u(t_i|\boldsymbol{z_i})\Big] \tag{A.3}$$

where $f_u(t_i)$ and $S_u(t_i)$ have the following expressions,

$$f_u(t_i) = h_{u0}(t_i)\frac{e^{\boldsymbol{\beta}'\boldsymbol{x_i}}}{\pi_i E[h_{u0}(t)]}\exp\left\{-H_{u0}(t_i)\frac{e^{\boldsymbol{\beta}'\boldsymbol{x_i}}}{\pi_i E[h_{u0}(t)]}\right\} \tag{A.4}$$

and

$$S_u(t_i) = \exp\left\{-H_{u0}(t_i)\frac{e^{\boldsymbol{\beta}'\boldsymbol{x_i}}}{\pi_i E[h_{u0}(t)]}\right\} \tag{A.5}$$

Peng and Dear (2000) demonstrated that using the EM algorithm for the standard cure rate model could separate the π_i and uncured distribution in the M step, thus makes the estimation much more convenient compared with the MLE method. However, this is not true for the marginal mean hazard rate model. According to above Equations A.3, A.4 and A.5, π_i appears in both the $f_u(t_i)$ and $S_u(t_i)$, then π_i cannot be fitted separately from the distribution of the uncured subpopulation in the M step. In other words, maximizing Equation A.3 would not be easier than maximizing likelihood function directly. Therefore, it is not beneficial to use the EM algorithm to estimate the marginal mean hazard rate model compared with using the MLE method.